

INTERNATIONAL TRACTS IN
**COMPUTER SCIENCE AND TECHNOLOGY
AND THEIR APPLICATION**

General Editors : N. METROPOLIS—Chicago
E. PIORE—New York
S. ULAM—Los Alamos
assisted by an International Honorary Editorial Advisory Board

VOLUME 2

SELF-ORGANIZING SYSTEMS

Already published in this series:

A PRACTICAL MANUAL ON THE MONTE CARLO METHOD FOR
RANDOM WALK PROBLEMS. By E. D. CASHWELL and
C. J. EVERETT, 1959.

SELF-ORGANIZING SYSTEMS

PROCEEDINGS OF AN
INTERDISCIPLINARY CONFERENCE

5 AND 6 MAY, 1959

Co-sponsored by

INFORMATION SYSTEMS BRANCH
OF OFFICE OF NAVAL RESEARCH

and

ARMOUR RESEARCH FOUNDATION
OF ILLINOIS INSTITUTE OF TECHNOLOGY

Editors

MARSHALL C. YOVITS

Office of Naval Research
CONFERENCE CHAIRMAN

SCOTT CAMERON

Armour Research Foundation
CONFERENCE SECRETARY

SYMPOSIUM PUBLICATIONS DIVISION

PERGAMON PRESS

OXFORD · LONDON · NEW YORK · PARIS

1960

PERGAMON PRESS LTD.
Headington Hill Hall, Oxford
4 & 5 Fitzroy Square, London W.1.

PERGAMON PRESS INC.
122 East 55th Street, New York 22, N.Y.
1404 New York Avenue N.W., Washington, D.C.

PERGAMON PRESS S.A.R.L.
24 Rue des Écoles, Paris Ve

PERGAMON PRESS G.m.b.H.
Kaiserstrasse 75, Frankfurt-am-Main

003.7

C74s

1959

cop. 3

Copyright

©

1960

PERGAMON PRESS INC.

LIBRARY OF CONGRESS CARD NUMBER: 60-12574

8 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100

Printed in Great Britain by Bell & Bain, Ltd., Glasgow

PREFACE

DURING the last decade or so there have been many great and important advances in the capabilities of information processing equipments and techniques. For the most part these have been based on the concept of the fixed stored program computer. This type of computer has shown itself able to solve large classes of problems which can first be put into some sort of numerical form and for which a program in machine language can be written.

Improvements in the speed, size, and sophistication of existing machines by factors of ten or a hundred appear to be near at hand, and a certain amount of research is presently directed toward the millimicrosecond or "ultimate" stored program computer. These advances will certainly open the way to making many more classes of problems conveniently tractable by digital computers.

However, it appears that certain types of problems, mostly those involving inherently non-numerical types of information, can be solved efficiently only with the use of machines exhibiting a high degree of learning or self-organizing capability. Examples of problems of this type include automatic print reading, speech recognition, pattern recognition, automatic language translation, information retrieval, and control of large and complex systems. Efficient solutions to problems of these types will probably require some combination of a fixed stored program computer and a self-organizing machine.

In recent months, it had become evident that interest in this field of cognitive systems was growing quite rapidly. A number of new groups had been formed to do research which would lead toward an understanding of these self-organizing systems. On the one hand the psychologist, the embryologist, the neurophysiologist and others involved in the life sciences were attempting to understand the self-organizing properties of biological systems, while mathematicians, engineers, and physical scientists were attempting

v

DEC 1 '78

ROBT LIBRARY
CARNEGIE-MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213

to design artificial systems which could exhibit self-organizing properties.

Accordingly, the Information Systems Branch of the Office of Naval Research together with Armour Research Foundation, decided to sponsor a conference enabling the workers in the many disciplines involved to meet together to discuss their research activities and to explore common problems, mutual interests, and similar directions of research. This volume comprises the Proceedings of that Conference, including the prepared papers and the associated discussion. The name, Interdisciplinary Conference on Self-Organizing Systems, was chosen in order to emphasize deliberately the broad nature of the research necessary to pursue these endeavors.

The papers for the program were chosen entirely by invitation of the Conference Committee and were intended to be representative of the appropriate research activities in progress by workers in the various disciplines. Most of the speakers invited had already achieved reputations within their respective fields and were known rather widely within the scientific community. Fourteen formal papers were presented, the authors of which provided almost equal representation of the fields of Biology, Engineering, Mathematics, and Psychology. In addition, an after-dinner address was presented at the Conference banquet by Dr. A. M. Uttley of the National Physical Laboratory, Teddington, England. The welcoming address was delivered by Dr. H. A. Leedy, Director of the Armour Research Foundation, and the opening address of the Conference was delivered by Dr. F. J. Weyl, Research Director of the Office of Naval Research. Almost 400 people, divided among the various scientific, engineering, and social-scientific disciplines, attended the Conference.

There are, of course, in addition to those who participated in the Conference, many other research workers of equally great reputation who are contributing significantly to an understanding of self-organizing systems. The Committee deeply regrets the inability to include these people on the program. To the many people who kindly offered to present papers at the Conference, the Committee gives its sincere thanks as well as its apologies for the inability to include these contributions. The Conference could easily have been extended several more days with contributed papers, most of which would have been of very high quality. Perhaps it would have been

wiser to plan the Conference for more than the two days which were actually used. It was the Committee's feeling, however, that significantly more of the important interested people would be able to attend a two-day meeting than one of longer duration and that the fundamental purposes of the Conference would be best accomplished by the shorter meeting.

Those paper that were presented appeared to fall naturally into four different interdisciplinary Groups. These were, with the authors of the papers in the order presented:

- I. *Perception of the Environment*
Farley; von Foerster; Estes; Rosenblatt.
- II. *Effects of Environmental Feedback*
Auerbach; Goldman; Bishop.
- III. *Learning in Finite Automata*
Newell, Shaw, Simon; Milner; Minsky; Campbell.
- IV. *Structure of Self-Organizing Systems*
Pask; McCulloch; Burks.

The Committee wishes to thank each of these authors and co-authors, as well as Drs. Leedy, Weyl and Uttley, for participating in the Conference and helping to make it a success. Their assistance is greatly appreciated.

The papers in Groups I and II were presented during the first day under the Chairmanship of Dr. Otto M. Schmitt, Departments of Zoology and Physics, University of Minnesota, Minneapolis, Minnesota. The papers in Groups III and IV were presented during the second day under the Chairmanship of Dr. John McCarthy, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts. The Committee wishes to express its thanks to Dr. Schmitt and Dr. McCarthy for the excellent way in which they handled the presentations and the subsequent discussions. They were instrumental in bringing out many interesting points that would not otherwise have arisen. The success of the Conference was due in large part to these Chairmen.

Each of the papers was scheduled for thirty minutes, with ten minutes allotted for questions. At the end of each day a panel, made up of the authors who spoke that day and the Chairman for

the day, held an open discussion, with questions and comments being welcomed from the audience. It was apparent in the editing that many of the comments in the open discussions were directed at specific papers which had been presented that day and were therefore accordingly placed after the appropriate papers in the written proceedings in the interest of clarity for the reader.

All of the papers presented at the Conference are included in these Proceedings in the order presented, each followed by the appropriate discussion as explained above, with the exception of the paper entitled "Progress on the Advice Taker" by Dr. Marvin Minsky, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts. Because of other commitments it was not possible for Dr. Minsky to submit his manuscript prior to publication of the Proceedings. It is expected that the paper will eventually be submitted to one of the scientific journals for publication.

The Committee which planned this Conference was made up of Marshall C. Yovits, Office of Naval Research, Chairman; Scott Cameron, Armour Research Foundation, Secretary; Albert R. Dawe, Office of Naval Research; Gordon D. Goldstein, Office of Naval Research; Harold Kantner, Armour Research Foundation; and Maynard Shelly, Office of Naval Research.

For the Committee,

MARSHALL C. YOVITS, Chairman
Head, Information Systems Branch
Office of Naval Research

CONTENTS

	PAGE
Preface...by M. C. YOVITS	v
Welcome Address...by H. A. LEEDY	1
Opening Address...by F. J. WEYL	3
Self-Organizing Models for Learned Perception...by B. G. FARLEY	7
On Self-Organizing Systems and Their Environments ...by H. VON FOERSTER	31
Statistical Models for Recall and Recognition of Stimulus Patterns by Human Observers...by W. K. ESTES	51
Perceptual Generalization over Transformation Groups ...by F. ROSENBLATT	63
The Organization and Reorganization of Embryonic Cells ...by R. AUERBACH	101
Further Consideration of Cybernetic Aspects of Homeostasis ...by S. GOLDMAN	108
Feedback Through the Environment as an Analog of Brain Functioning...by G. H. BISHOP	122
FIRST DAY'S PANEL DISCUSSION	147
A Variety of Intelligent Learning in a General Problem Solver ...by A. NEWELL, J. C. SHAW and H. A. SIMON	153
Learning in Neural Systems...by P. M. MILNER	190
Blind Variation and Selective Survival as a General Strategy in Knowledge-Processes...by D. T. CAMPBELL	205
The Natural History of Networks...by G. PASK	232
The Reliability of Biological Systems...by W. S. McCULLOCH	262
Computation, Behavior, and Structure in Fixed and Growing Automata...by A. W. BURKS	282
SECOND DAY'S PANEL DISCUSSION	312
The Mechanization of Thought Processes...by A. M. UTTLEY	319

1

2

3

4

5

6

7

WELCOME ADDRESS

H. A. LEEDY

*Director, Armour Research Foundation,
Illinois Institute of Technology*

I AM pleased to have the opportunity to welcome you to this Conference on behalf of the Armour Research Foundation and to say that we hope you will find your visit to Chicago an enjoyable one. We also hope that you will find the events of the next two days both stimulating and rewarding. The Armour Research Foundation has been pleased to have the opportunity to work with the Office of Naval Research in the organization of this rather unique Conference. While the concept of an interdisciplinary conference is certainly not a new one, the universality of the problem which this Conference has assembled to discuss is attested to not only by the makeup of the formal program and the group in attendance, but also by the extensive correspondence we have received from scientists in such diverse fields as psychology, linguistics, neurophysiology, embryology, information theory, biology, psychiatry, mathematics, cosmology, and the social sciences as well as those concerned with the development of information systems. While to a large extent this interest can be explained in terms of man's natural preoccupation with himself and his apparent ability to create islands of decreasing entropy, it is our feeling that it also derives from an increasing conviction that a coherent theory of organization is essential to the solution of the truly overwhelming information processing problems which confront our modern civilization. It is our hope that this Conference will represent at least a small step in this direction. Although the interdisciplinary makeup of the Conference will undoubtedly create certain problems in communication, we feel that the benefits to be derived from the resolution of these problems more than compensates for the effort expended.

At this time I should like to introduce to you Dr. F. Joachim Weyl, Research Director of the Office of Naval Research, who will present the opening address.

OPENING TALK

Dr. F. JOACHIM WEYL

*Research Director, Office of Naval Research
Washington, D.C.*

GOOD morning, self-organizing systems. I am indeed very happy to find the Office of Naval Research joining with the Armour Research Foundation in organizing this Conference on what I personally consider an exceedingly important topic, and doing this at such a well chosen time, as the attendance indicates.

The choice of the time is particularly significant in my personal life, too. For the last nine months the Department of Defense of the United States of America has been in the throes of an organizational effort which shows reasonably clearly that we are still a long way from understanding what makes a self-organizing system.

Now, joking aside, let us not completely dismiss the military establishment of a country as large as the United States. It furnishes an example of one major class of real systems of the type that we must learn to understand in order to get to the bottom of the story that will be told during the next few days. The other two, flanking the social organization, are respectively the computer systems and the biological organisms. I have the feeling that we will rapidly find growing up a vocabulary in the field that draws freely on the preformed vocabularies from those three classes of organizations. Each of them has one characteristic aspect I feel, to teach us, and brings with it a typical class of problems whose understanding and, if you will pardon my harkening back to my own professional past, mathematical transluminations will be the important job of this field.

From the area of computers we will in the long run draw our essential understanding of the element of memory that is absolutely and inevitably present in what you might designate in the future as self-organizing systems. You might go so far, and I have done it,

as to say that a computer is nothing but a means for a memory to get from one state to another.

I would say that the biological organism will probably have to be relied on to furnish the insight into the second of these major basic elements. I think the biologists have learned to call it differentiation. In any system that will evolve it is quite clearly necessary that some events take place that will split from a mass of homogeneous, of alike elements, one group specializing in one direction as distinct from another.

The third of these basic elements I would like to call to your attention can be studied best, I would say, in social organizations. Let me call it, for the purpose here, subordination, or if you wish, the executive function. It probably presents itself most purely and most accessibly when we are dealing with these large social organizations.

All three of them have a certain number of elements in common. I think you will all agree with me that to-day the concept of organization that is emerging is essentially a communication and information theoretical one. As such, you will find inevitably two things entering. On the one hand, the entire mathematical apparatus of signal to noise problems, statistical in nature, making use of such apparatuses as that of testing hypotheses, decision theory, and so forth. The other physical complex of ideas that enters depends on the nature of time and its characteristic irreversibility in any physical system.

When dealing with memory in all of its aspects, you realize that for evolutionary processes to take place you need the equivalent of what the geneticists have called mutations, essentially random events. You need these supplemented by a natural selection process that filters out from these random fluctuations about a static equilibrium, those that lie to the right, as it were, on the evolutionarily upper directed branch, from those that lie to the left along evolutionarily downward directed branches.

In the problem of differentiation you will find problems of the same nature inevitably occurring. Starting with the homogeneous group of elements that are destined for differentiation, some initial triggering mechanism is needed to push one group in one direction as distinguished from another in another direction. In other words, environment containing noise has to be relied on to furnish the triggering mechanism on which the long-term selection rule will operate.

The last remark I wanted to call to your attention lies in the following direction. A great deal has already been thought and said about the characteristic operation of evolutionary systems and a nomenclature is growing up that I first encountered in a paper by Hans Bremermann where he speaks of the Eigen-model of its environment that such a system carries, systematically working at its improvement.

The titles of the papers following in the next two days use different words to describe similar situations. Now in this storage process of Eigen-models of the environment one tends to stress always, in the area of biological systems in particular, the similarity between the genetic apparatus and the central nervous system as systems that operate in this fashion. I would like briefly to point out also that there is a profound complementarity between them which you will seek quickly if I step across the border to the class of social organizations as an example. Anyone who has concerned himself with social organizations as I have during the last few months (in the attempt to see how the Office of Naval Research ought to look in the future) will realize that he has to deal with two types of organizational patterns, you might call it the formal organization and the informal organization. In a sense, the formal organization I would like to parallel to the genetic apparatus in the biological system. It contains, as it were, the *a priori* pattern that is given for an organizational apparatus. The informal organization, on the other hand, I would like to parallel to the central nervous system which within that pattern establishes a kind of loosely drawn, continually changing and quite locally adaptable pattern of hierarchically altering, informational traffic. So that in a certain way, and very speculatively, I would like to call to your attention the possible fruitfulness of seeing systematically how two such different types of organizations, the predetermined, predrawn, blueprint type of pattern on the one hand, and the fluid, adaptive type of organization on the other interact and influence each other in the creation of such systems.

You could not, I hope did not, expect from me more than such indications of where I think fruitful problems lie. I have no solutions, but I am an eager learner and shall listen to what follows in order to see where I discern the beginnings of answers. I hope that as the papers of the next few days roll by, you will really make

them sound as if you are dealing with some of the most important problems. I for one will be with you, because I do believe that they are.

Thank you.

SELF-ORGANIZING MODELS FOR LEARNED PERCEPTION*

BELMONT G. FARLEY†

M.I.T., Lincoln Laboratory, Lexington, Mass.

Abstract—A framework of ideas is suggested for models of systems which automatically organize themselves to classify environmental inputs into recognizable percepts or "patterns." The models operate by computing "properties" of environmental inputs and comparing the results with stored classes of properties to select percepts. Property-classes may be formed from existing lists of properties by operating on the environmental input with suitable rules, and the computation of additional properties may be organized also.

It is suggested that such models, especially on account of their non-linear character, should be able to perform many of the functions of learned perception as observed in living organisms. They should also prove useful for engineering and scientific purposes. Neurophysiological realization of the models seems possible.

Much investigation of the behavior of such models with various rules in various environments is necessary, however, to verify these suggestions.

INTRODUCTION

EVERY living organism must be able to classify the input stimuli presented by its environment in such a way as to ensure its survival. Simple organisms depend to a large extent on relatively specific predetermined classification procedures, while the more sophisticated possess in large measure the ability to organize their classification procedures to take account of increasingly complex features of their environment. This self-organizing ability is called "learned perception."

Much information about perception and its learning has been accumulated, and many theories proposed to account for the facts or various aspects of them. Allport (1955) has reviewed many of

* The work reported in this paper was performed by Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology with the joint support of the U.S. Army, Navy, and Air Force.

† Staff member, Massachusetts Institute of Technology, Lincoln Laboratory.

these theories. This paper represents an attempt to contribute to the class of ideas for theories about learned perception which try to account for at least the main stumbling-blocks of perception and its self-organization, and at the same time leave room for neurophysiological realization in plausible ways. Needless to say, there remain many gaps in the formulation, and the intent is to construct a framework or "apparatus" of ideas to suggest directions for further work, rather than present a "theory" in any rigorous sense.

A short discussion of previous work will be given first.

Perhaps the first attempts to link perception and physiology were made by the gestalt school of psychologists, who also discovered so many facts of perception which provide stumbling blocks to theory. These attempts, particularly by Lashley and Kohler, are reviewed by Allport (1955). They postulated fields in the cortex, which provided interactions by forces and interference patterns. These attempts, although interesting and provocative, were not made very definite, and were not related to physiology at the neuron level.

Pitts and McCulloch (1947) constructed a "mathematico-neurological" model to account for perception of "universals" of form, but did not suggest a mechanism for learning. Walter (1953) constructed specific models for conditioning, but applied them only to this simple case. Shimbelt (1950) and others also discussed some simple models of neural learning. Reference should also be made to the work of Frankel (1955).

D. O. Hebb, in his book *The Organization of Behavior*, was perhaps the first to consider the psychological facts of perception in considerable detail and relate them to a fairly specific neurophysiological model. Although his neurophysiological model contained serious defects, his discussion of the problems in perceptual theory remains provocative. We will assume the reader is familiar with at least the first three chapters. His work was the stimulus for several efforts to investigate neural models, among them the first paper of the present author with W. A. Clark (1954). Since this work is related to the possible neurophysiological realization of the property-class scheme to be presented later, a brief description will be given here.

The system simulated was a randomly interconnected network of nonlinear elements, each element having a threshold for incoming

excitation, below which no action occurs, and above which the element "fires." When an element fires, its threshold immediately rises to infinity, and then, after a short refractory period, falls exponentially back toward its quiescent value. Furthermore, a short time after firing, an element transmits excitation to all other elements to which it is connected. The effectiveness of the excitation thus transmitted to a succeeding element is determined by a property of the particular connection known as its "weight." In general, there will be several incoming connections at any element, each having its individual weight as shown in Fig. 1. At the instant of

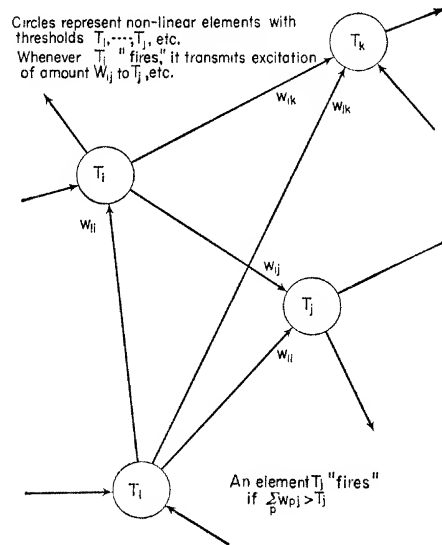


FIG. 1. Network of neuron-like elements.

transmission the appropriate weight is added to any excitation already present at the succeeding cell. Thereafter the excitation decays exponentially to zero. Thus the model contained both "spatial" and "temporal" summation. If at any time this excitation exceeds the threshold of the succeeding element, that

element performs its own firing cycle and transmits its own excitations. Note that the element simulated was analog, not simply a logical function.

The network was activated and an output obtained in the following way. The net was divided arbitrarily into two sets, designated as input and output sets. The output set was further subdivided in two, and an output was defined at any instant by the difference in the number of elements fired in the two subsets during the instant. This arrangement might be termed a "push-pull" output.

The input set was also subdivided into two subsets, and two fixed input patterns were provided, designated as p_1 and p_2 . Input p_1 consisted in adding a large excitation periodically into all the input elements of one subset, but doing nothing to the other subset. Input p_2 reversed the role of the subsets. In this way output activity characteristic of the input pattern was obtained.

It was now found possible to provide a modifier acting upon parameters of the net so as to gradually reorganize it to obtain output activity of a previously specified characteristic, namely, so that patterns p_1 and p_2 always drive the output in previously specified directions. In our experiments, p_1 was made to drive the output in a negative direction and p_2 to drive it positive.

This desired organization of the net was accomplished by varying the weights in the following way. Examination is made of the change in output at every instant. If a change in a favorable direction occurs (e.g. negative change in case p_1 is the input pattern), then all weights which just previously participated in firing an element are increased. If, on the other hand, the change was unfavorable, those weights are decreased. It is important to note that there is no detailed examination of the internal activity of the net. As a result, some of the weights may be altered in the wrong direction at any given time. However, as our results show, in the long run a favorable result occurs, when p_1 and p_2 are presented alternately.

Thus the system organized itself to distinguish between two distinct input patterns.

Rochester (1956), using a somewhat simpler neuron model, obtained similar results, while attempting to realize Hebb's model more closely.

When a net has been trained to discriminate between two patterns, it becomes of interest to investigate how it will classify, without

further training, patterns which it has not seen before. In Clark and Farley (1955) this was done. It was found that, statistically, a pattern tended to be classified like one of the original training patterns if it had input elements in common with it—the more

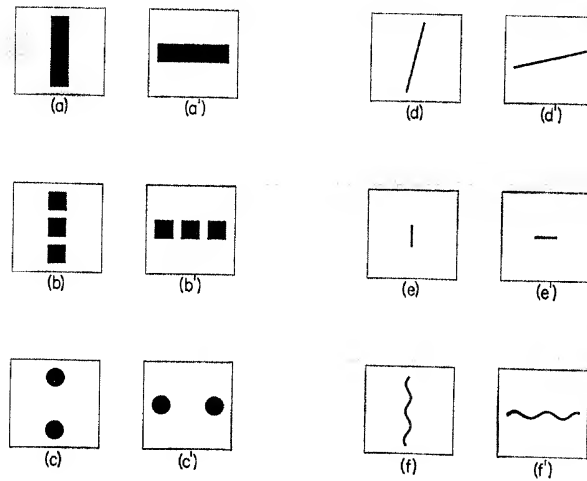


FIG. 2. Horizontal-vertical discrimination patterns.

elements in common, the closer the association. Using Fig. 2 it was pointed out that, after training for discrimination of horizontal and vertical bars, this simple type of "generalization" would suffice to discriminate correctly the other horizontal and vertical figures shown without any further training, assuming automatic centering. The reason for this generalization is simply that each horizontal test figure has more elements in common with the original horizontal training figure than elements in common with the original vertical training figure, and vice versa for the vertical case.

During testing and training, it was shown that a considerable amount of noise could be added to the figure elements without affecting the results, and it was pointed out that this can have the effect of widening such an "overlap" generalization even further.

More recently, Rosenblatt (1958) has considered a random network organized somewhat differently. He found that these nets

exhibited a generalization essentially the same as the one described above, and has shown how this effect may be utilized to distinguish different classes of patterns by training with a suitably representative subset of each class rather than having to train the net with all specimens of each class. This result is possible because almost any reasonable test figure will overlap some figures of its own training class more than those of the other class, if the training sets were suitably "representative." Thus, a class can be memorized using nearly-overlapping figures for training, instead of all figures of the class. Note, however, that the "overlap" generalization takes no account of shape or topological properties of figures. One way of doing so is to structure the nets, and Rosenblatt has also investigated some such cases.

It is interesting to note that Hebb (p. 48), suggested that memorization utilizing overlapping sets might make it feasible to learn all examples of one kind, such as horizontal lines. It should be pointed out that such a memorization may be thought of as learning a primitive "property," in this case "horizontalness." Means of using such properties will be suggested later.

It should be mentioned here that Taylor (1955) and Uttley (1956) have considered nets and modifiers belonging to this same class of ideas.

Some Difficult Problems in Perception

The work so far discussed indicates that means are available for learning certain types of classes (or properties) of an environment by memorizing enough representative or typical specimens of the class. It seems hardly likely, however, that the more complex stimuli of the environment are learned by such simple schemes.

For one thing, many experimental facts of perception are very hard to explain on such bases. No attempt will be made here to give a systematic account of perception, since this would be a task beyond the competence of the writer, but some problems of perception which appear to be among the hardest to deal with will be mentioned. More detailed accounts will be found in Hebb, Allport, Gibson (1950), Koffka (1935), Metzgar (1953), and other treatises on the psychology of perception.

As has been suggested above, one of the main problems in constructing models of learned perception is to provide suitable means

of "generalization" from particular stimuli to new stimuli never before encountered. Such mechanisms are necessary unless the environment to be dealt with is so limited in extent and so amenable to control that every possible stimulus can be presented and memorized individually. Such a simple case never occurs in nature, and rarely occurs even in useful artificial situations. Usually, the stimuli vary greatly, sometimes in ways which are not predictable. Simple examples are the change in the retinal image of objects when viewed from different angles of regard or distances. However, objects may even be recognized when partly obscured, or actually partly changed, as when a friend wears a new outfit. Indeed, it may be doubted whether a retinal image is ever repeated exactly.

It should be noted that it is important to distinguish between generalizations which result from simply ignoring differences, and those which result from predictions about the environment—from sample to class. Both kinds undoubtedly occur. The "overlap" generalization of the nets is of the first kind, insofar as figure recognition is concerned.

Another way of talking about generalization is to use the term "similarity." Establishing suitable generalizations may be thought of as equivalent to establishing a criterion of similarity, in the simplest case perhaps a space of stimuli with a "distance" measure of similarity. Classifications of stimuli would then be made on the basis of their similarities. The nets discussed above provide an example of a primitive similarity measure—two inputs are similar according to the number of elements they have in common.

A number of problems are posed by the principles of perception found experimentally by the gestalt school. A few of these will be mentioned in an oversimplified way.

One is "closure and good continuation." We have already seen how the random nets may effect generalizations of this kind, from full line to dotted line, etc.

A second principle is "wholeness." The whole perceptual field determines the percept, often expressed by the statement "the whole is greater than the sum of all its parts." An example of this is a well-known illusion. Two parallel lines no longer appear parallel when intersected by a "pencil" of lines crossing at angles. Instead they appear bowed inward or outward depending on how the intersecting set of lines is drawn.

A third principle invokes "field-forces" to explain why, if one dot is somewhat out of line in an otherwise perfect dotted circle, it usually appears in or closer to the circle of dots than it really is. A "force" is said to pull the dot back. Another example is the fact that a parallelogram appears more rectangular than it really is. A related phenomena is that certain figures or configurations ("good" figures) play preferred roles, as shown by experiments in tachistoscopic and impoverished perception. In such cases, when the subject does not have enough time, or light, or has poor seeing conditions for some other reason, there is a distinct tendency to prefer certain configurations or shapes, and as the seeing becomes clearer, the preferred figures change gradually to more realistic ones. See, for example, Metzgar (1953), Chapter 7.

Any account of perception must also be able to explain the "constancies." A dinner plate appears to be a dinner plate, and it appears round no matter what the angle of regard. This is the essence of perception.

Another point is that much perception seems related closely to motor functions. Hebb discusses the part eye movement may play. Gibson (1950), and Sperry (1952), have discussions of this point, and Allport (1955) has reviewed the matter. A close connection between motor activity and perception is also suggested by the fact that transfer of learning takes place between senses. For example, a shape learned visually can be immediately recognized tactually, or, in case its image is too big to fall all at once on the retina, it can be recognized by moving the eyes or head, or even the whole body. It would seem that space perception must be closely tied up with motor activity. After all, much perception is a prelude to movement.

A final consideration is that attention, "set," motivation, and context play very important parts. They are what Hebb calls "non-sensory" factors in perception.

Any notions looking toward models for perception must have room for plausible inclusion of at least the above factors and be neurophysiologically plausible as well. The construction of a complete model or theory of learned perception satisfying even the above boundary conditions would be a formidable task, and no such claim is made for what follows. It would be useful, however, in lieu of a theory, to have a conceptual framework for learned

perception with plausibility arguments which indicate some possibility of including the behaviors discussed. Such a framework will at least suggest means for further testing and synthesis.

Self-Organization of Property Classes

The idea of a percept as an association, class, or "bundle" of properties probably goes back at least as far as Aristotle and has reappeared in one form or another many times since then. See, for example, Gibson, p. 222. In fact, listing of properties has always been one of the main occupations of the descriptive sciences. However, the idea does not seem to have been put forward in a form definite enough to indicate how it might eventually evolve into a model of perception.

There would seem to be sound reasons from both biological and psychological considerations for attacking perception from this standpoint.

On the biological side, Tinbergen (1951) has performed experiments with fish in order to discover just what visual impressions are necessary to release characteristic types of behavior. Similar experiments have been performed with birds. These experiments appear to show that certain simple combinations of simple properties suffice in many cases, and that other aspects of the object tend to be ignored as irrelevant.

Now consider a child and his concept of the object "dog." To him, with a limited experience, let us say, of only one dog, the concept "dog" can be said to consist of a bundle of "properties" of color (perhaps brownish), hairyness (all over), legs (several), noise (bark, growl), teeth (sharp, many), odor (doggy), audible class name ("dog"), movement (head first, rapid), size (medium), head (where eyes and teeth are), tail (opposite head), etc. On his first contact with a dog, only a few of these properties may be noted, perhaps only a brownish blur and a bark, but with further familiarization other properties such as those mentioned above and many more gradually become grouped together, until eventually subclasses of "dog" based again on further properties are distinguishable, and furthermore, dogs are distinguished from other animals by reason of differing lists of properties possessed in differing amounts.

Thus perhaps a model of the classification process can be constructed as follows. We consider a data-processing system which

has the ability to compute measures of a given list of properties from any inputs fed it over a certain period of time. These properties may include weight, color, temperature, or other physical characteristics and may also include counts of corners, statements about shape or type of movement, or relationships between other properties, etc. For example, if shape is a property measurable by our system, the statement "the shape is triangular" may be considered as fixing the "value" (triangular) of the variable property "shape." It can be seen that every statement which can be made about an input may be considered as determining a value of a property variable.

Other properties useful in visual perception may be the angles of lines, curvature of lines or boundaries, and the gradients of texture emphasized as so important by Gibson.

For auditory perception, properties such as frequency and intensity are certainly measured. Others useful for speech recognition, might be "formant" frequencies and their rates of change, and noise timing, for example.

As indicated above, every object, or (more generally) percept, may be defined by a list of the properties it possesses and the observed ranges (or better, the probability distributions) of the values of each property. For the time being, it will be considered that the list of properties measurable by our system is fixed. We can see how such a system could recognize and respond to classes of input data by means of a chart which is schematized in Fig. 3. Each measurable property is listed horizontally as p_1, p_2, p_3 , etc., and its values plotted vertically. Each different class of properties makes up a percept, several of which are shown as c_1, c_2 , etc. In each case a distribution of property values is indicated. In order to determine to which percept a given input belongs, it is only necessary to measure its properties and see which class c_i it fits best.

One interesting advantage of this model is the vast number of classes c_i which can be distinguished by means of a few properties. For example, if one hundred measurable properties are available, and if they average ten distinguishable distributions of value each, then the number of distinguishable classes is of the order of 10^{100} . A great deal of the "storage" for the class recognition is thus accounted for by the computation of property measurements, which are used over and over again.

Now the question arises as to how the property classes are compiled, since this is the essence of the self-organizing process.

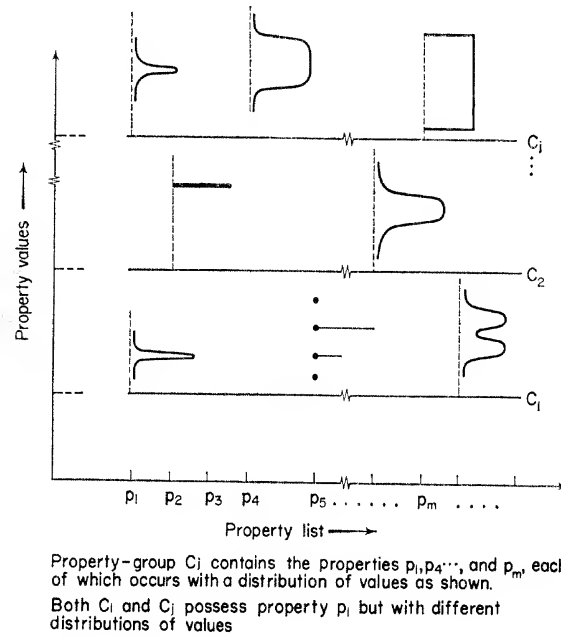


FIG. 3. Scheme showing property-classes and value distributions.

In order to carry out such a compiling process, some criteria must be used to determine which properties are to be grouped together. Several possible rules of procedure come to mind. In the first place, the properties constituting a percept usually appear repeatedly together, both in space and time. Therefore, the system must examine the frequency of simultaneous or nearly simultaneous occurrence of properties. Rules must then operate to consider properties with a number of associative repetitions to constitute a group, perhaps tentatively at first, becoming gradually more certain as repetitions increase, or a repetition frequency threshold may be established, above which the properties are grouped.

Furthermore, property-classes are more likely to be meaningful if one or more property values are constant or relatively so over a given period of time, or under particular conditions. For example,

when examining an object in the hands, its temperature, weight, and odor, if any, remain constant although many of its visual properties may change radically as it is turned about. Thus, procedures should be in force which group properties according as they occur in combination with other "decision" properties whose values remain roughly constant. Some of these "decision" properties may, of course, be given more weight than others. Sometimes overriding weight may be given to one property, as when the system is "told" that a figure belongs to a particular class. Note that such a process may lead to a relatively sudden acquisition of a new class, when the frequency criterion for a particular sub-group of properties is exceeded.

We have discussed property-grouping rules based on frequency of occurrence, contiguity in space, continuity in time, and presence of certain key "decision" properties. The expectation is that with a suitable set of initial rules, a self-organizing system, when exposed to a reasonable input environment of properties, will gradually build up property-classes like those in Fig. 3, and that the system will then be able to recognize a percept when its property-class shows a sufficient correlation with a listed one. Naturally, the exact nature of the grouping rules is not known, and the above discussion represents a plausible guess based on biological and psychological considerations of existing self-organizing systems.

There are other aspects of this model of the classification process whose investigation should prove very interesting. For one, it can be seen that a primitive symbolization of a percept can come about through the frequent occurrence of a particular sub-group of properties in conjunction with the property-class of that percept. In this way, a "name" repeated often enough nearly simultaneously with the appearance of a significant property-class can eventually become connected with the class and may be considered as a name-symbol of the corresponding percept.

Furthermore, if means are available to the system for the inter-comparison of the property-classes c , sub-groups of properties or values may be found which are common to many percepts. Such sub-groups have the character of new, abstract, properties or "concepts" which may be used in future organization. For example, if it were found that many useful percepts of the environment were objects having three straight boundaries and three

corners, the more abstract concept "triangle" might result from the process described above.

More should be said about the nature of the rules which determine the "best fitting" class when comparison of classes is being made with incoming data. The exact nature of good rules of this type remain to be investigated, but it seems clear that the correlation should involve a threshold which, if exceeded, would indicate the choice of the appropriate class. A very simple rule would be majority vote of properties common to class and input, perhaps with some properties weighted more heavily than others. Note that correlation rules and thresholds need not be constant, and those property-classes under active consideration may also change with external conditions. Such changes could account for the psychological phenomena of "set" or attention, and motivation. "Context" can exert its influence in the same way—correlation thresholds for example, can change with time or space "surroundings."

The classification scheme described above can be pictured in a multidimensional space of the properties. Then the probability distributions of values of properties in a class may be thought of as forming a "blob" about a point representing the class. Separate classes are then separated blobs in the space and the distance between blobs becomes a similarity measure. An input is then a single point in the space, and in the simplest case would be identified with the closest blob. (Actually, the space needed may not be quite as simple as this, since as we have seen, some properties may be more heavily weighted than others, and the correlation criterion may produce a rather peculiar "distance" function, so that reduction to a space with at least a triangular inequality everywhere may be somewhat complicated.) We have mentioned the correlation or "distance" thresholds which would result in "neighborhoods" near blobs.

A physical realization of such a system might be expected to possess very non-linear static and dynamic characteristics, having for example, a "stable point" in the vicinity of each class blob, with nearby trajectories leading to the stable point. (There are a number of treatises on non-linear systems. See, for example, Cunningham, 1958.) Figure 4 shows a schematic diagram of such a space with stable points.

The non-linear action may then result in a kind of "locking-in" of input to class. A rough one-dimensional analogy in engineering terms would be the lock-in action of the AFC in a radio receiver.

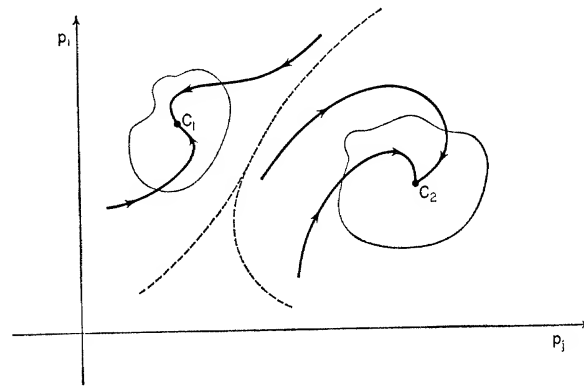


FIG. 4. Stable points and trajectories leading to classes in "property-space."

In the multidimensional case, on a strong signal (many sensory input properties giving a strong correlation with some property-class) the action would be expected to be quick and secure. In fact, it might be so strong as to lock on some property-class in spite of the fact that certain properties or their values were in conflict with that class. As a result, a percept would have been attained which did not altogether correspond with the real sensory stimulus. Depending on the exact action of the system upon "attaining a percept" (reaching a stable point), the conflicting properties might be ignored altogether, or their values might be "pulled in" toward the values to be expected in the class corresponding to the stable point. If two whole classes conflicted, an oscillation between classes could occur, but only one would be selected at any one time.

We may now compare the results of such behavior with some of the facts of perception mentioned earlier.

In the first place, it is clear that in a non-linear system of this kind, two figures do not necessarily superpose perceptually. The percept arrived at will, in fact, be determined by the "whole" figure, or situation. Furthermore, we have seen that oscillation

between two percepts may occur, if they conflict. This is most likely to occur if distinguishing properties are missing, as in unstable illusions. Two such percepts, however, are never seen simultaneously.

It must be pointed out that the distributions of properties may not be independent. For example, slant to line-of-sight, as measured by textural gradient is correlated with projected shape. Ordinarily, when the system selects a class, all the properties will fall in their respective distributions, and they will all agree with one another as to dependence. In this case, which is the usual one in perception, constancy of object holds, and its properties are all normal. However, if one or a few property-values are out of expected distribution, we have seen that the system action may be to "pull" them toward their expected or preferred values. Such action may explain those effects for which the gestalt school were led to postulate "forces" of visual organization. We may then expect some figures to "appear" more like preferred ones in some respects, or for compromises in angles or distances to be made.

In addition, these preferred figures—preferred by reason of their lower thresholds or large "lock-in" volumes—would be expected to be most likely to be perceived under conditions of impoverished viewing. Impoverished viewing corresponds here to "seeing" conditions with comparatively few properties available for comparison.

It can be seen from these examples that it is at least conceivable that many perceptual experiments in the gestalt realm can be simulated by a property-class model of the kind described, provided that suitable rules of class formation and correlation can be found. Since it is quite feasible to simulate such a system on a general purpose computer, its behavior under various assumptions can be examined. This simulation could be carried out rather abstractly, but perhaps the best way would be to try out such a system on an actual environment provided by an engineering or scientific problem. It seems clear that even if the property-class type of self-organizing system fails to provide the framework of a perceptual theory, it should provide a quite practical means of building a system of learned pattern-recognition, in those engineering or scientific problems for which it is possible to think of some plausible properties beforehand.

An illustration of a partial application of these ideas to a scientific problem is shown in Fig. 5 (Farley *et al.*, 1957), which shows a portion of a human electroencephalogram. The top of the step

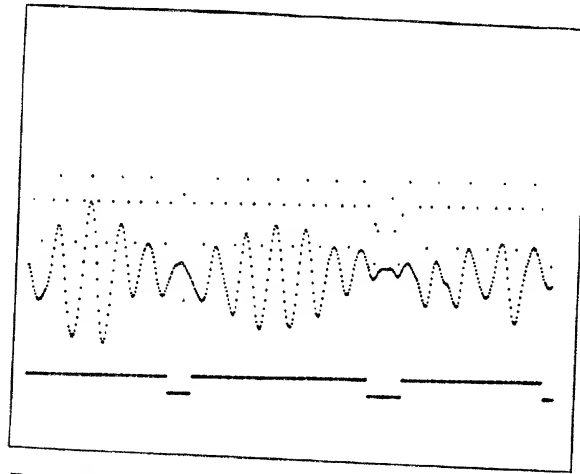


FIG. 5. Detection of "rhythmic-burst" pattern in an electroencephalogram.

function indicates the presence of a "rhythmic burst" of activity as determined by a computer program. The pertinent point here is that the program considers a "rhythmic burst" as a conjunction of two simple properties of each cycle—satisfaction of an amplitude threshold criterion, and satisfaction of a frequency range criterion. That is, wherever a "clump" of cycles having these two properties occur together the pattern of "burst activity" is recognized. It is clear that this is a simple case of a system of property-class perception of the type we have been discussing. Only two properties have been used, and their occurrence together was expected beforehand in this case as a means of detecting the bursts. But there seems to be every reason to expect that with a suitable list of properties, the wave could be explored automatically, noting which properties occur together frequently, and thus organizing property-classes in the way we have described.

Since the basic problem of experimental science is to determine what repeatabilities (percepts) exist in the data, procedures such as these may be expected to play an important part in scientific data

processing, particularly in the behavioral sciences which produce so much complex data.

Organization of Properties

So far, it has been assumed that the property-list used for organizing property-classes was fixed. It is of interest also to consider how the list came to exist and how it might be extended.

In the first place the detection and measurement of some properties may be built into the system from the start. This appears to be true of many properties in lower animals. It seems likely that the ability to measure some primitive properties is built into man also, although this has been a controversial issue for years. A possible example is that property which enables the infant to direct his eyes toward a light.

Another possibility already suggested is that certain properties may be learned simply by memorizing all "essentially different" instances of the property, and thus forming a property-class which expresses the new property. This can be done by using a "decision property" and is thus a special case of the property-class organization scheme already described. Memorization thus effected, is an example of the possibility of building up new properties from more primitive ones.

Another method of obtaining new properties is suggested by the above discussion. As each new property-class is organized the fact that an input belongs to it may be considered a property and added to the list. Combinations of classes (classes of classes) may be treated in the same way.

Such considerations lead naturally to properties of properties as new properties. The use of this idea may also be illustrated by the electroencephalographic study. In Fig. 6 is plotted the number of "rhythmic burst" patterns in a standard interval versus the amplitude threshold for four different subjects, showing patterns characteristic of each subject. (We omit discussion here of the statistical fluctuations of each subject to simplify our point.) Here we have an example of the case in which a property (number per time) of property-classes (rhythmic bursts) produces a useful new percept characteristic of individual subjects.

Another means of forming new properties has already been mentioned—the formation of "abstractions" consisting of sub-groups

of properties common to many property-classes. The property of containing such a sub-group may be added to the list. Use of those abstractions as "decision" properties is an important way of making predictive generalizations.

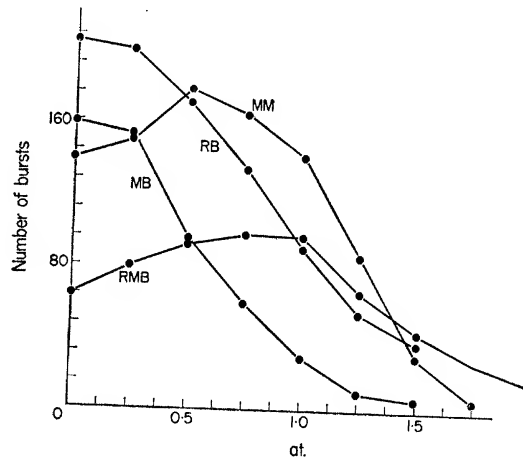


FIG. 6. Number of bursts vs. amplitude threshold for several subjects.

Incidentally, it should be clear that the system will contain most of the desirable features from a purely logical standpoint—hierarchies of properties, classes, and means of symbolization, including the possibility of gradually building up more complex systems.

Neurophysiological Models

We have yet to investigate whether our model can be plausibly represented by the functions of the nervous system. Unfortunately, the paucity of information about the functional organization of the nervous system makes such discussion necessarily speculative, and the mere statement that a function does not conflict with what is known may well be true, but also trivial. However, it is important to have some ideas in mind as a guide to further development and testing, so a few somewhat disconnected suggestions will be made.

First we may consider correlations of inputs with stored property-groups by neurophysiological means. It may be remembered that

in the testing of all possible input patterns to a net which had been trained to discriminate between two input sets p_1 and p_2 , an analog output was derived which was used to classify the inputs. This output may in fact be considered as a function measuring the correlation of the test input with p_1 and p_2 . Thus, the network of the experiment may actually be regarded as a simple prototype of a property-class correlation system for classification. Indeed, the original patterns were also learned by using a combination of the key-decision property (the "favorable direction" feedback), and frequency of repetition (in this case repetition of use of certain connections). We have also seen how random networks of this type can memorize classes of inputs and therefore build up properties.

The exact mechanisms tried so far may not be actual ones used by the nervous system since others can be imagined which should also work, but it seems plausible that this general type of memorization may play some part in neurophysiological learning but no experimental evidence for this is known. Each property-dimension learned by a random net might be expected to have "overlap" generalization and this may be advantageous. Since the basic requirements appear plausible, we may consider that a classification model could be realized. Much more work, both experimental and theoretical, will be required to make this more definite.

One way might be mentioned that properties measured successively in time can be grouped by repetition if the time constants of the nervous tissues are suitable. In the case of eye fixations, for example, the fact that a property exists could be indicated by activity in a particular set of neurons corresponding to that property. This activity might last for some time. Another property may be registered similarly during the next eye fixation, and if the two sets of neurons interact in any area, weights of connections to the same cell might grow with repetition and a new set of neurons eventually activated corresponding to the frequent close time coincidence of the two properties. In this way property-classes could be built up.

Certain properties naturally measured by the nervous system may be related rather directly to the motor apparatus. The skeleto-muscular system is a natural set of standard axes, which is reflected in the anatomical organization of the cortex. Since the sensory system is likewise represented in the cortex, it could be imagined

that the two sets are related in such a way as to result in the natural measurement of properties relative to the body axes, such as left, right, up-down, and relative position such as "up-to-the-left," etc. It is interesting that a description in these terms is independent of translation, size, and small rotations, which agrees with experiments on many species of animal. Transfer of recognition from visual to tactile could be accounted for by some such mechanism, and possibly some aspects of space perception as well.

Characteristic activity in a neuron network as a result of gradients of texture as has already been mentioned seems a likely possibility, which would result in the measurement of properties whose importance in visual perception have been emphasized by Gibson. (See particularly his Chapter 5.)

One of the difficult stumbling-blocks of neurophysiological theories of perception has been that large chunks of some parts of the cortex can often be removed without affecting perception noticeably. It seems that the property-class concept can overcome this obstacle by invoking an excess of properties in a property-class over a minimum set, and a redundancy of the representation of properties and property-classes. The activity representing a given property or property-class need not be confined to a particular part of the cortex. Indeed, in some respects, the more widespread the activity corresponding to a property, the better, since it will then be enabled to interact with many others to form classes.

It should be noted in this connection that some properties appear to be measured in subcortical structures. This is indicated, for example, by Butler *et al.* (1957), where it is shown that ablation of the auditory cortex in cat leaves an ability to discriminate changes in frequency.

SUMMARY

It appears that the property-class model described represents a plausible model for many aspects of learned perception. Means are suggested for organizing percepts (property-classes) from more primitive properties, and behaviors to be expected of the system, which is highly non-linear, resemble some perceptual behaviors in biology and psychology. However, exact rules for input correlation with classes, and for organizing classes and properties are not known and must be investigated.

Arguments have also been presented which are intended to make plausible the idea that a self-organizing property-class system can be constructed in neurophysiological form. However, no experimental evidence exists, and much work obviously remains to be done to see if such notions hold up.

In any case, it is felt that a property-class model should be of use in engineering and scientific pattern recognition.

REFERENCES

- ALLPORT, F. H. (1955). *Theories of Perception and the Concept of Structure*, John Wiley, New York.
- BEURLE, R. L. (1956). Properties of a mass of cells capable of regenerating pulses, *Phil. Trans. Roy. Soc. London B* **240**, No. 669, 55.
- BUTLER, R. A., DIAMOND, T. T., and NEFF, W. D. (1957). Role of auditory cortex in discrimination of changes in frequency, *J. Neurophysiol.* **20**, 108.
- CLARK, W. A. and FARLEY, B. G. (1955). Generalization of pattern recognition in a self-organizing system, *Proceedings of the Western Joint Computer Conference*, Los Angeles, Cal.
- CUNNINGHAM, W. J. (1956). *Introduction to Nonlinear Analysis*, McGraw-Hill, New York.
- FARLEY, B. G. and CLARK, W. A. (1954). Simulation of self-organizing systems by digital computer, *Trans. IRE*, vol. PGIT-4, 76-84.
- FARLEY, B. G., FRISHKOPF, L. S., CLARK, W. A., and GILMORE, J. T. (1957). Computer techniques for the study of patterns in the electroencephalogram. Technical Report 165, Lincoln Laboratory, M.I.T. (Presented at the Michigan Symposium on Pattern Recognition, 1957.)
- FRANKEL, S. (1955). On the design of automata and the interpretation of cerebral behavior, *Psychometrika* **20**, No. 2.
- GIBSON, J. J. (1950). *The Perception of the Visual World*, Houghton Mifflin, Cambridge.
- GRANIT, R. (1955). *Receptors and Sensory Perception*, Yale University Press.
- HEBB, D. O. (1949). *The Organization of Behavior*, John Wiley, New York.
- KOFFKA, K. (1935). *Principles of Gestalt Psychology*, Harcourt, Brace, New York.
- LASHLEY, K. S. (1942). The problem of cerebral organization in behavior. Vol. VII of *Biological Symposia*, p. 302, Cattell, London.
- METZGAR, W. (1953). *Gesetze des Sehens*, Kramer, Frankfurt-am-Main.
- PITTS, W. and MCCULLOCH, W. S. (1947). How we know universals, the perception of auditory and visual forms, *Bull. Math. Biophysics* **9**, 127.
- ROCHESTER, N., HOLLAND, J. H., HAIBT, L. H., and DUDA, W. L. (1956). Tests on a cell assembly theory of the action of the brain, using a large digital computer, vol. PGIT-2, No. 3, 80.
- ROSENBLATT, F. (1958). The perceptron, Cornell Aeronautical Lab., Buffalo, N.Y. Rept. VG-1196-G-1.
- ROSENBLATT, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain, *Psychol. Rev.* **65**, 386.
- ROSENBLITH, W. A. (1957). Relations between auditory psychophysics and auditory electrophysiology. *Trans. N.Y. Acad. Sci. Ser. II*, **19**, No. 7, 650.

- SELFRIDGE, O. (1958). Pandemonium, a paradigm for learning, *Proc. of Symp. on Mechanization of Thought Processes*, Teddington, England.
- SHIMBEL, A. (1950). Contributions to the mathematical biophysics of the central nervous system, with special reference to learning, *Bull. Math. Biophys.* **12**, 241.
- SHOLL, D. A. (1956). *The Organization of the Cerebral Cortex*, John Wiley, New York.
- SPERRY, R. W. (1952). The mind-brain problem, *Am. Scientist* **40**, No. 2.
- TAYLOR, W. K. (1955). The electrical simulation of some nervous system functional activities, *Proc. of the Third London Symp. on Information Theory*.
- TAYLOR, W. K. (1959). Pattern recognition by means of automatic analogue apparatus. *Proc. Inst. Elect. Engrs.* **106B**, No. 26, March, 198-209.
- TINBERGEN, N. (1951). *Study of Instinct*, Oxford University Press.
- UTTLEY, A. M. (1956). Conditional probability machines and conditioned reflexes, in *Automata Studies* (ed. by C. E. SHANNON and J. MCCARTHY), Princeton University Press. Further Refs. at the end of this paper.
- WALTER, W. G. (1953). *The Living Brain*, Norton.
- YOUNG, J. Z. (1957). *Doubt and Certainty in Science: A biologist's reflections on the brain*, Oxford University Press.

DISCUSSION

CHAIRMAN SCHMIDT: If I may take the privilege of starting this discussion I would like to ask Dr. Farley whether he would care to comment on the relatively slight attention that has been paid to the topological and geometrical configuration of neuronal organizations in making up these models for mathematical analysis of their functions? That is, the cortex, for example, is quite elaborately structured and has symmetry of an interesting kind which has entered rather little into the discussions that I have heard of this matter. I wonder if you would care to comment on this?

FARLEY: Well, I think it is just a reflection of the primitive state of the science as yet. That is, we just haven't gotten around to it. I think everybody has it in mind.

TOWSTER: How is the learning time affected by the number of elements in your nerve nets?

FARLEY: Well, we didn't have enough samples to make a very good statement about that. I suppose you refer to real time rather than computer time. Obviously, computer time is a lot longer for larger nets because you have to process so many more things.

If I were asked to hazard a guess I would suspect it wouldn't be very much different in real time for large nets for the simple types of processes which we have discussed. It wouldn't take much longer. It might even be shorter because the larger the net the better the statistics.

ROSENBLATT: I would like to back up Bel's comment on that with some observations on our own systems which are organized in a somewhat similar fashion in many cases.

You will recall, I think, that in the slide this morning which showed the spontaneous learning experiment, we are dealing with an infinite perceptron. That still required runs of some thousands of stimuli, before we got the perfect dichotomization of the two classes. We run into a limitation which is not a function of the size of the system. In small systems there is an additional time increment which can be due to the smallness of the system. But once we get

into a fair-sized system in simple discrimination problems, then the fundamental time limitation in this type of generalization, or our contiguity generalization, comes about because we require a large enough sample of stimuli to cover the retinal field effectively, and there will be a finite waiting time until we achieve such a sample. This is quite fundamental.

KANTER (*Armour Research Foundation*): I might start out some additional discussion by suggesting that it seemed to me in the latter part of your discussion, the property space was rather hazy and your opening remarks related to some network approach that you have been engaged in. I would like to have your comments on the representation of property space by a network.

FARLEY: Well, of course, this is necessarily quite vague at the present time, so all I can do is make a couple of very vague suggestions. Hebb, as you know, had a scheme which involved neurophysiological analogs. While this neurophysiological scheme had certain serious defects it suggested a good many things to us. The work of Uttley, for example, has shown certain nonlinearities in systems of neurons in which, as you get the neurons over a certain threshold, activity will continue until some disturbance is given. Many neurons in the cortex have recurrent collaterals which suggest self-excitation. These are just vague speculations at the present time.

KANTER: Well, I didn't exactly have reference to a physiological network, but rather, if we consider the property space as a mathematical object, have you in mind, or might you suggest, a representation of the property space as a mathematical object comprised of the linear graph?

FARLEY: I haven't done anything detailed along those lines. I merely want to point out it may be rather difficult. I mentioned that the property space was a first approximation. The reason I say that is that this is a very peculiar space. It may not even have a distance function unless you cook up something rather complicated because of the peculiarities of the correlation function itself. For example, majority rule would result in a rather odd distance measure. Furthermore, the property space tends to "waver" around a good bit, I think, because of the questions of context and motivation and so forth. I don't know off hand whether this first approximation would tell you anything more than your intuition tells you now.

BRETZ (*University of Michigan*): In your diagram of the neural model there weren't any cycles shown. In the experiments that you conducted on learning did you use cycles in the neurons themselves or just in the feedback mechanism that changed the threshold?

FARLEY: There were cycles in the net itself. The figure was just an element of the net intended to represent a piece of it. It was actually constructed at random just by making up a connection matrix and throwing "ones" into the matrix at random so you had the possibility of circulating activity.

As a matter of fact, this causes trouble with self-excitation. We got rid of it in a way, which I won't go into now, but you can read it in the paper. This is, incidentally, one of the problems that you always run into when you are talking about neural networks. The question is, why doesn't everybody have epilepsy all the time? (Laughter).

MCCARTHY (*M.I.T.*): You mentioned this idea which goes back to Aristotle, that an object is a collection of properties. It seems to me that this idea is very dubious in that even when you introduce so simple a thing as the integers, which you do for counting purposes, then you are getting away from this idea and introducing instead a collection of objects or a collection of entities which, in order to be distinguished require more than a finite set of properties.

FARLEY: Well, I think there we would have to digress into questions of the philosophy of mathematics. I, myself, am very dubious as to what extent you can say that infinite classes exist. You can talk about them, but does this mean they exist? And in any case, I can't really believe it has anything to do with scientific questions, because any scientific theory can be represented as closely as you please by a finite system if it is large enough.

McCARTHY: Suppose we consider some entities like algebraic expressions or texts which are composed of many letters or circuit diagrams. I guess circuit diagrams are actually the best example. We recognize circuit diagrams ourselves by their recursive structure. That is the way they are built out of their parts rather than properties of the diagrams of a whole.

FARLEY: From my point of view, these recursive things are properties in themselves.

McCARTHY: Yes, but you can't learn to recognize one of them and then learn to recognize some fixed set of them and then expect to be able to apply this to all circuit diagrams, at least it doesn't seem to me you can.

FARLEY: I wasn't suggesting learning a fixed set. Means were suggested for acquiring new sets. Of course, I don't want to leave the impression that you can necessarily explain every possible fact of perception by the scheme I have talked about. On the other hand, it seems to me even in mathematics, all classes are known by their properties.

ON SELF-ORGANIZING SYSTEMS AND THEIR ENVIRONMENTS*

H. von FOERSTER

*Department of Electrical Engineering
University of Illinois, Urbana, Illinois*

I AM somewhat hesitant to make the introductory remarks of my presentation, because I am afraid I may hurt the feelings of those who so generously sponsored this conference on self-organizing systems. On the other hand, I believe, I may have a suggestion on how to answer Dr. Weyl's question which he asked in his pertinent and thought-provoking introduction: "What makes a self-organizing system?" Thus, I hope you will forgive me if I open my paper by presenting the following thesis: ["There are no such things as self-organizing systems !"] ✓

In the face of the title of this conference I have to give a rather strong proof of this thesis, a task which may not be at all too difficult, if there is not a secret purpose behind this meeting to promote a conspiracy to dispose of the Second Law of Thermodynamics. I shall now prove the non-existence of self-organizing systems by *reductio ad absurdum* of the assumption that there is such a thing as a self-organizing system.

Assume a finite universe, U_0 , as small or as large as you wish (see Fig. 1a), which is enclosed in an adiabatic shell which separates this finite universe from any "meta-universe" in which it may be immersed. Assume, furthermore, that in this universe, U_0 , there is a closed surface which divides this universe into two mutually exclusive parts: the one part is completely occupied with a self-organizing system S_0 , while the other part we may call the environment E_0 of this self-organizing system: $S_0 \& E_0 = U_0$.

I may add that it is irrelevant whether we have our self-organizing system inside or outside the closed surface. However, in Fig. 1 the

* Supported by the Information Systems Branch of the Office of Naval Research under Contract Nonr. 1834 (21).

system is assumed to occupy the interior of the dividing surface. Undoubtedly, if this self-organizing system is permitted to do its

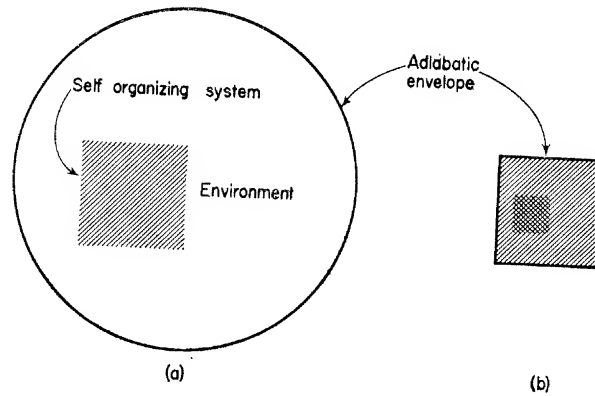


FIG. 1.

job of organizing itself for a little while, its entropy must have decreased during this time:

$$\frac{\delta S_s}{\delta t} < 0,$$

otherwise we would not call it a self-organizing system, but just a mechanical $\delta S_s/\delta t = 0$, or a thermodynamical $\delta S_s/\delta t > 0$ system. In order to accomplish this, the entropy in the remaining part of our finite universe, i.e. the entropy in the environment must have increased

$$\frac{\delta S_E}{\delta t} > 0,$$

otherwise the Second Law of Thermodynamics is violated. If now some of the processes which contributed to the decrease of entropy of the system are irreversible we will find the entropy of the universe U_0 at a higher level than before our system started to organize itself, hence the state of the universe will be more disorganized than before $\delta S_U/\delta t > 0$, in other words, the activity of the system was a disorganizing one, and we may justly call such a system a "dis-organizing system."

However, it may be argued that it is unfair to the system to make it responsible for changes in the whole universe and that this apparent inconsistency came about by not only paying attention to the system proper but also including into the consideration the environment of the system. By drawing too large an adiabatic envelope one may include processes not at all relevant to this argument. All right then, let us have the adiabatic envelope coincide with the closed surface which previously separated the system from its environment (Fig. 1b). This step will not only invalidate the above argument, but will also enable me to show that if one assumes that this envelope contains the self-organizing system proper, this system turns out to be not only just a disorganizing system but even a self-disorganizing system.

It is clear from my previous example with the large envelope, that here too—if irreversible processes should occur—the entropy of the system now within the envelope must increase, hence, as time goes on, the system would disorganize itself, although in certain regions the entropy may indeed have decreased. One may now insist that we should have wrapped our envelope just around this region, since it appears to be the proper self-organizing part of our system. But again, I could employ that same argument as before, only to a smaller region, and so we could go on for ever, until our would-be self-organizing system has vanished into the eternal happy hunting grounds of the infinitesimal.

In spite of this suggested proof of the non-existence of self-organizing systems, I propose to continue the use of the term “self-organizing system,” whilst being aware of the fact that this term becomes meaningless, unless the system is in close contact with an environment, *which possesses available energy and order*, and with which our system is in a state of perpetual interaction, such that it somehow manages to “live” on the expenses of this environment.

Although I shall not go into the details of the interesting discussion of the energy flow from the environment into the system and out again, I may briefly mention the two different schools of thought associated with this problem, namely, the one which considers energy flow and signal flow as a strongly linked, single-channel affair (i.e. the message carries also the food, or, if you wish, signal and food are synonymous) while the other viewpoint carefully

separates these two, although there exists in this theory a significant interdependence between signal flow and energy availability.

I confess that I do belong to the latter school of thought and I am particularly happy that later in this meeting Mr. Pask, in his paper *The Natural History of Networks*,⁽²⁾ will make this point of view much clearer than I will ever be able to do.

What interests me particularly at this moment is not so much the energy from the environment which is digested by the system, but its utilization of environmental order. In other words, the question I would like to answer is: "How much order can our system assimilate from its environment, if any at all?"

Before tackling this question, I have to take two more hurdles, both of which represent problems concerned with the environment. Since you have undoubtedly observed that in my philosophy about self-organizing systems the environment of such systems is a *conditio sine qua non* I am first of all obliged to show in which sense we may talk about the existence of such an environment. Second, I have to show that, if there exists such an environment, it must possess structure.

The first problem I am going to eliminate is perhaps one of the oldest philosophical problems with which mankind has had to live. This problem arises when we, men, consider ourselves to be self-organizing systems. We may insist that introspection does not permit us to decide whether the world as we see it is "real," or just a phantasmagory, a dream, an illusion of our fancy. A decision in this dilemma is in so far pertinent to my discussion, since—if the latter alternative should hold true—my original thesis asserting the nonsensicality of the conception of an isolated self-organizing system would pitifully collapse.

I shall now proceed to show the reality of the world as we see it, by *reductio ad absurdum* of the thesis: this world is only in our imagination and the only reality is the imagining "I".

Thanks to the artistic assistance of Mr. Pask who so beautifully illustrated this and some of my later assertions,* it will be easy for me to develop my argument.

Assume for the moment that I am the successful business man with the bowler hat in Fig. 2, and I insist that I am the sole reality,

*Figures 2, 5 and 6

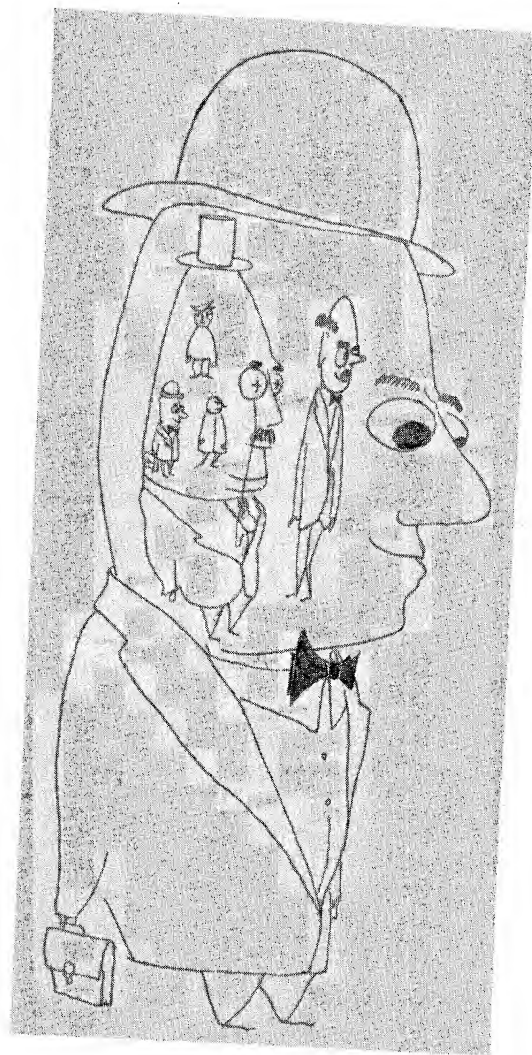
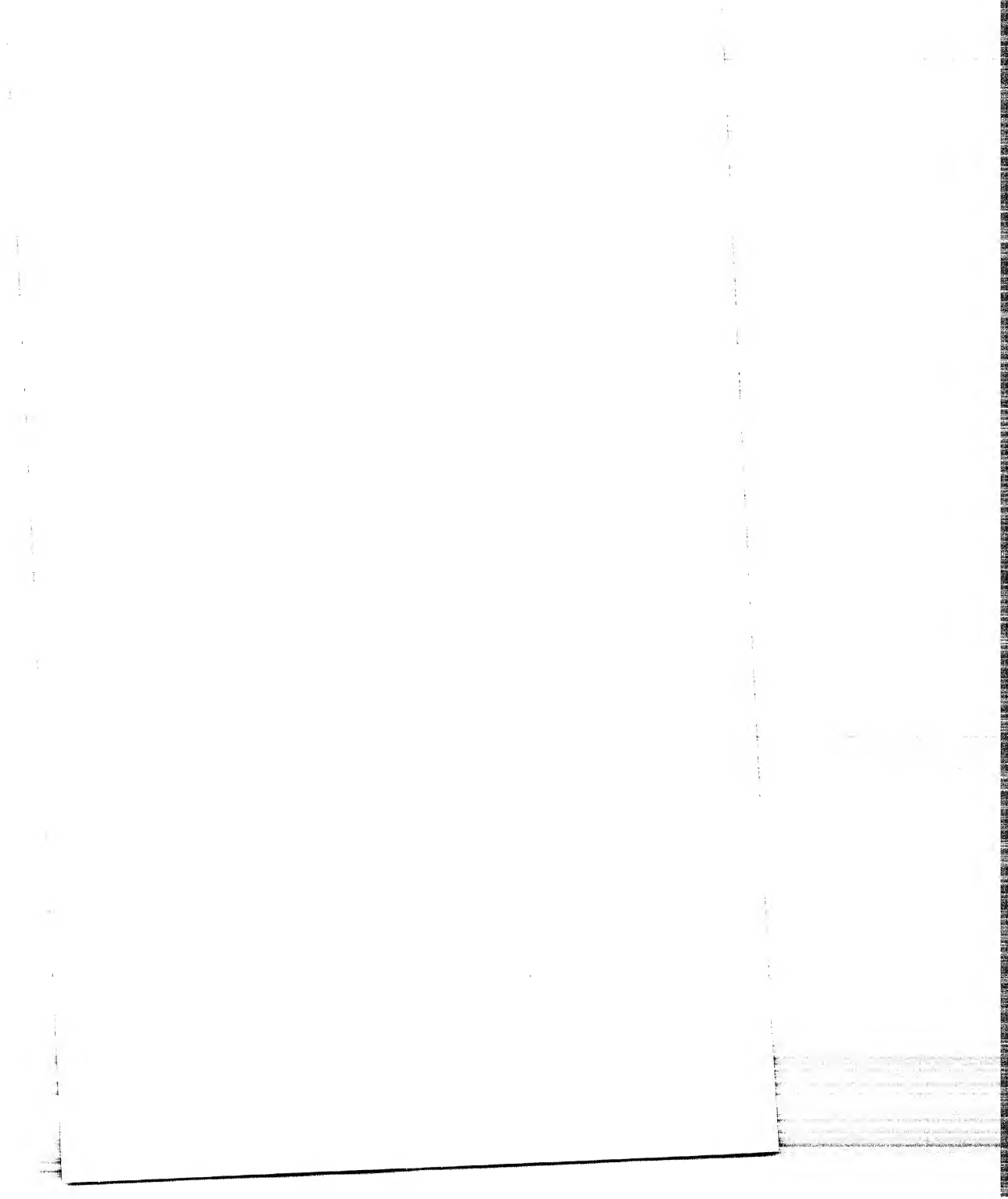


FIG. 2.



while everything else appears only in my imagination. I cannot deny that in my imagination there will appear people, scientists, other successful businessmen, etc., as for instance in this conference. Since I find these apparitions in many respects similar to myself, I have to grant them the privilege that they themselves may insist that they are the sole reality and everything else is only a concoction of their imagination. On the other hand, they cannot deny that their fantasies will be populated by people—and one of them may be I, with bowler hat and everything!

With this we have closed the circle of our contradiction: If I assume that I am the sole reality, it turns out that I am the imagination of somebody else, who in turn assumes that *he* is the sole reality. Of course, this paradox is easily resolved, by postulating the reality of the world in which we happily thrive.

Having re-established reality, it may be interesting to note that reality appears as a consistent reference frame for at least two observers. This becomes particularly transparent, if it is realized that my "proof" was exactly modeled after the "Principle of Relativity," which roughly states that, if a hypothesis which is applicable to a set of objects holds for one object and it holds for another object, then it holds for both objects simultaneously, the hypothesis is acceptable for all objects of the set. Written in terms of symbolic logic, we have:

$$(Ex) [H(a) \& H(x) \rightarrow H(a + x)] \rightarrow (x) H(x) \quad (1)$$

Copernicus could have used this argument to his advantage, by pointing out that if we insist on a geocentric system, $[H(a)]$, the Venusians, e.g. could insist on a venucentric system $[(Hx)]$. But since we cannot be both, center and epicycloid at the same time $[H(a + x)]$, something must be wrong with a planetocentric system.

However, one should not overlook that the above expression, $\mathcal{R}(H)$ is not a tautology, hence it must be a meaningful statement.* What it does, is to establish a way in which we may talk about the existence of an environment.

* This was observed by Wittgenstein,⁽⁶⁾ although he applied this consideration to the principle of mathematical induction. However, the close relation between the induction and the relativity principle seems to be quite evident. I would even venture to say that the principle of mathematical induction is the relativity principle in number theory.

Before I can return to my original question of how much order a self-organizing system may assimilate from its environment, I have to show that there is some structure in our environment. This can be done very easily indeed, by pointing out that we are obviously not yet in the dreadful state of Boltzmann's "Heat-Death." Hence, presently still the entropy increases, which means that there must be some order—at least now—otherwise we could not lose it.

Let me briefly summarize the points I have made until now:

- (1) By a self-organizing system I mean that part of a system that eats energy and order from its environment.
- (2) There is a reality of the environment in a sense suggested by the acceptance of the principle of relativity.
- (3) The environment has structure.

Let us now turn to our self-organizing systems. What we expect is that the systems are increasing their internal order. In order to describe this process, first, it would be nice if we would be able to define what we mean by "internal," and second, if we would have some measure of order.

The first problem arises whenever we have to deal with systems which do not come wrapped in a skin. In such cases, it is up to us to define the closed boundary of our system. But this may cause some trouble, because, if we specify a certain region in space as being intuitively the proper place to look for our self-organizing system, it may turn out that this region does not show self-organizing properties at all, and we are forced to make another choice, hoping for more luck this time. It is this kind of difficulty which is encountered, e.g., in connection with the problem of the "localization of functions" in the cerebral cortex.

Of course, we may turn the argument the other way around by saying that we define our boundary at any instant of time as being the envelope of that region in space which shows the desired increase in order. But here we run into some trouble again; because I do not know of any gadget which would indicate whether it is plugged into a self-*dis*organizing or self-organizing region, thus providing us with a sound operational definition.

Another difficulty may arise from the possibility that these self-organizing regions may not only constantly move in space and change in shape, they may appear and disappear spontaneously

here and there, requiring the "ordometer" not only to follow these all-elusive systems, but also to sense the location of their formation.

With this little digression I only wanted to point out that we have to be very cautious in applying the word "inside" in this context, because, even if the position of the observer has been stated, he may have a tough time saying what he sees.

Let us now turn to the other point I mentioned before, namely, trying to find an adequate measure of order. It is my personal feeling that we wish to describe by this term two states of affairs. First, we may wish to account for apparent relationships between elements of a set which would impose some constraints as to the possible arrangements of the elements of this system. As the organization of the system grows, more and more of these relations should become apparent. Second, it seems to me that order has a relative connotation, rather than an absolute one, namely, with respect to the maximum disorder the elements of the set may be able to display. This suggests that it would be convenient if the measure of order would assume values between zero and unity, accounting in the first case for maximum disorder and, in the second case, for maximum order. This eliminates the choice of "neg-entropy" for a measure of order, because neg-entropy always assumes finite values for systems being in complete disorder. However, what Shannon⁽³⁾ has defined as "redundancy" seems to be tailor-made for describing order as I like to think of it. Using Shannon's definition for redundancy we have:

$$R = 1 - \frac{H}{H_m} \quad (2)$$

whereby H/H_m is the ratio of the entropy H of an information source to the maximum value, H_m , it could have while still restricted to the same symbols. Shannon calls this ratio the "relative entropy." Clearly, this expression fulfills the requirements for a measure of order as I have listed them before. If the system is in its maximum disorder $H = H_m$, R becomes zero; while, if the elements of the system are arranged such that, given one element, the position of all other elements are determined, the entropy—or the degree of uncertainty—vanishes, and R becomes unity, indicating perfect order.

What we expect from a self-organizing system is, of course, that, given some initial value of order in the system, this order is going to increase as time goes on. With our expression (2) we can at once state the criterion for a system to be self-organizing, namely, that the rate of change of R should be positive:

$$\frac{\delta R}{\delta t} > 0 \quad (3)$$

Differentiating eq. (2) with respect to time and using the inequality (3) we have:

$$\frac{\delta R}{\delta t} = - \frac{H_m(\delta H/\delta t) - H(\delta H_m/\delta t)}{H_m^2} \quad (4)$$

Since $H_m^2 > 0$, under all conditions (unless we start out with systems which can only be thought of as being always in perfect order: $H_m = 0$), we find the condition for a system to be self-organizing expressed in terms of entropies:

$$H \frac{\delta H_m}{\delta t} > H_m \frac{\delta H}{\delta t} \quad (5)$$

In order to see the significance of this equation let me first briefly discuss two special cases, namely those, where in each case one of the two terms H , H_m is assumed to remain constant.

(a) $H_m = \text{const.}$

Let us first consider the case, where H_m , the maximum possible entropy of the system remains constant, because it is the case which is usually visualized when we talk about self-organizing systems. If H_m is supposed to be constant the time derivative of H_m vanishes, and we have from eq. (5):

$$\text{for } \frac{\delta H_m}{\delta t} = 0 \dots \dots \frac{\delta H}{\delta t} < 0 \quad (6)$$

This equation simply says that, when time goes on, the entropy of the system should decrease. We knew this already—but now we may ask, how can this be accomplished? Since the entropy of the system is dependent upon the probability distribution of the elements to be found in certain distinguishable states, it is clear that this probability distribution must change such that H is reduced. We

may visualize this, and how this can be accomplished, by paying attention to the factors which determine the probability distribution. One of these factors could be that our elements possess certain properties which would make it more or less likely that an element is found to be in a certain state. Assume, for instance, the state under consideration is "to be in a hole of a certain size." The probability of elements with sizes larger than the hole to be found in this state is clearly zero. Hence, if the elements are slowly blown up like little balloons, the probability distribution will constantly change. Another factor influencing the probability distribution could be that our elements possess some other properties which determine the conditional probabilities of an element to be found in certain states, given the state of other elements in this system. Again, a change in these conditional probabilities will change the probability distribution, hence the entropy of the system. Since all these changes take place internally I'm going to make an "internal demon" responsible for these changes. He is the one, e.g. being busy blowing up the little balloons and thus changing the probability distribution, or shifting conditional probabilities by establishing ties between elements such that H is going to decrease. Since we have some familiarity with the task of this demon, I shall leave him for a moment and turn now to another one, by discussing the second special case I mentioned before, namely, where H is supposed to remain constant.

(b) $H = \text{const.}$

If the entropy of the system is supposed to remain constant, its time derivative will vanish and we will have from eq. (5)

$$\text{for } \frac{\delta H}{\delta t} = 0 \dots \dots \frac{\delta H_m}{\delta t} > 0 \quad (7)$$

Thus, we obtain the peculiar result that, according to our previous definition of order, we may have a self-organizing system before us, if its possible maximum disorder is increasing. At first glance, it seems that to achieve this may turn out to be a rather trivial affair, because one can easily imagine simple processes where this condition is fulfilled. Take as a simple example a system composed of N elements which are capable of assuming certain observable states. In most cases a probability distribution for the number of elements

D

in these states can be worked out such that H is maximized and an expression for H_m is obtained. Due to the fact that entropy (or, amount of information) is linked with the logarithm of the probabilities, it is not too difficult to show that expressions for H_m usually follow the general form*:

$$H_m = C_1 + C_2 \log_2 N.$$

This suggests immediately a way of increasing H_m , namely, by just increasing the number of elements constituting the system; in other words a system that grows by incorporating new elements will increase its maximum entropy and, since this fulfills the criterion for a system to be self-organizing (eq. 7), we must, by all fairness, recognize this system as a member of the distinguished family of self-organizing systems.

It may be argued that if just adding elements to a system makes this a self-organizing system, pouring sand into a bucket would make the bucket a self-organizing system. Somehow—to put it mildly—this does not seem to comply with our intuitive esteem for members of our distinguished family. And rightly so, because this argument ignores the premise under which this statement was derived, namely, that during the process of adding new elements to the system the entropy H of the system is to be kept constant. In the case of the bucket full of sand, this might be a ticklish task, which may conceivably be accomplished, e.g. by placing the newly admitted particles precisely in the same order with respect to some distinguishable states, say position, direction, etc. as those present at the instant of admission of the newcomers. Clearly, this task of increasing H_m by keeping H constant asks for superhuman skills and thus we may employ another demon whom I shall call the “external demon,” and whose business it is to admit to the system only those elements, the state of which complies with the conditions of, at least, constant internal entropy. As you certainly have noticed, this demon is a close relative of Maxwell’s demon, only that to-day these fellows don’t come as good as they used to come, because before 1927⁽⁴⁾ they could watch an arbitrary small hole through which the newcomer had to pass and could test with arbitrary high accuracy his momentum. To-day, however, demons watching

* See also Appendix.

closely a given hole would be unable to make a reliable momentum test, and vice versa. They are, alas, restricted by Heisenberg's uncertainty principle.

Having discussed the two special cases where in each case only one demon is at work while the other one is chained, I shall now briefly describe the general situation where both demons are free to move, thus turning to our general eq. (5) which expressed the criterion for a system to be self-organizing in terms of the two entropies H and H_m . For convenience this equation may be repeated here, indicating at the same time the assignments for the two demons D_i and D_e :

$$\begin{array}{ccccccc}
 H & \times & \frac{\delta H_m}{\delta t} & > & H_m & \times & \frac{\delta H}{\delta t} & (5) \\
 \uparrow & & \uparrow & & \uparrow & & \uparrow \\
 \text{Internal} & & \text{External} & & \text{External} & & \text{Internal} \\
 \text{demon's} & & \text{demon's} & & \text{demon's} & & \text{demon's} \\
 \text{results} & & \text{efforts} & & \text{results} & & \text{efforts}
 \end{array}$$

From this equation we can now easily see that, if the two demons are permitted to work together, they will have a disproportionately easier life compared to when they were forced to work alone. First, it is not necessary that D_i is always decreasing the instantaneous entropy H , or D_e is always increasing the maximum possible entropy H_m ; it is only necessary that the product of D_i 's results with D_e 's efforts is larger than the product of D_e 's results with D_i 's efforts. Second, if either H or H_m is large, D_e or D_i respectively can take it easy, because their efforts will be multiplied by the appropriate factors. This shows, in a relevant way, the interdependence of these demons. Because, if D_i was very busy in building up a large H , D_e can afford to be lazy, because his efforts will be multiplied by D_i 's results, and vice versa. On the other hand, if D_e remains lazy too long, D_i will have nothing to build on and his output will diminish, forcing D_e to resume his activity lest the system ceases to be a self-organizing system.

In addition to this entropic coupling of the two demons, there is also an energetic interaction between the two which is caused by the energy requirements of the internal demon who is supposed to accomplish the shifts in the probability distribution of the elements comprising the system. This requires some energy, as we may remember from our previous example, where somebody has to blow up the little balloons. Since this energy has been taken from the environment, it will affect the activities of the external demon who may be confronted with a problem when he attempts to supply the system with choice-entropy he must gather from an energetically depleted environment.

In concluding the brief exposition of my demonology, a simple diagram may illustrate the double linkage between the internal and the external demon which makes them entropically (H) and energetically (E) interdependent.

For anyone who wants to approach this subject from the point of view of a physicist, and who is conditioned to think in terms of thermodynamics and statistical mechanics, it is impossible not to refer to the beautiful little monograph by Erwin Schrodinger *What is Life*.⁽⁵⁾ Those of you who are familiar with this book may remember that Schrodinger admires particularly two remarkable features of living organisms. One is the incredible high order of the genes, the "hereditary code-scripts" as he calls them, and the other one is the marvelous stability of these organized units whose delicate structures remain almost untouched despite their exposure to thermal agitation by being immersed—e.g. in the case of mammals—into a thermostat, set to about 310°K.

In the course of his absorbing discussion, Schrodinger draws our attention to two different basic "mechanisms" by which orderly events can be produced: "The statistical mechanism which produces order from disorder and the ... [other] one producing 'order from order'."

While the former mechanism, the "order from disorder" principle is merely referring to "statistical laws" or, as Schrodinger puts it, to "the magnificent order of exact physical law coming forth from atomic and molecular disorder," the latter mechanism, the "order from order" principle is, again in his words: "the real clue to the understanding of life." Already earlier in his book Schrodinger develops this principle very clearly and states: "What

an organism feeds upon is negative entropy." I think my demons would agree with this, and I do too.

However, by reading recently through Schrodinger's booklet I wondered how it could happen that his keen eyes escaped what I would consider a "second clue" to the understanding of life, or—if it is fair to say—of self-organizing systems. Although the principle I have in mind may, at first glance, be mistaken for Schrodinger's "order from disorder" principle, it has in fact nothing in common with it. Hence, in order to stress the difference between the two, I shall call the principle I am going to introduce to you presently the "order from noise" principle. Thus, in my restaurant self-organizing systems do not only feed upon order, they will also find noise on the menu.

Let me briefly explain what I mean by saying that a self-organizing system feeds upon noise by using an almost trivial, but nevertheless amusing example.

Assume I get myself a large sheet of permanent magnetic material which is strongly magnetized perpendicular to the surface, and I cut from this sheet a large number of little squares (Fig. 3a). These

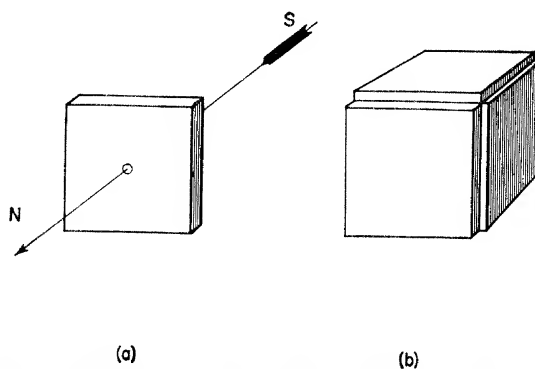


FIG. 3. (a) Magnetized square.
(b) Cube, family I.

little squares I glue to all the surfaces of small cubes made of light, unmagnetic material, having the same size as my squares (Fig. 3b). Depending upon the choice of which sides of the cubes have the

magnetic north pole pointing to the outside (Family I), one can produce precisely ten different families of cubes as indicated in Fig. 4.

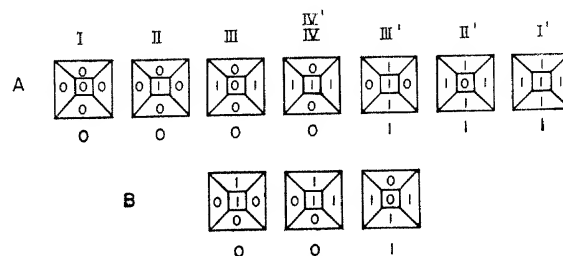


FIG. 4. Ten different families of cubes (see text).

Suppose now I take a large number of cubes, say, of family I, which is characterized by all sides having north poles pointing to the outside (or family I' with all south poles), put them into a large box which is also filled with tiny glass pebbles in order to make these cubes float under friction and start shaking this box. Certainly, nothing very striking is going to happen: since the cubes are all repelling each other, they will tend to distribute themselves in the available space such that none of them will come too close to its fellow-cube. If, by putting the cubes into the box, no particular ordering principle was observed, the entropy of the system will remain constant, or, at worst, increase a small amount.

In order to make this game a little more amusing, suppose now I collect a population of cubes where only half of the elements are again members belonging to family I (or I') while the other half are members of family II (or II') which is characterized by having only one side of different magnetism pointing to the outside. If this population is put into my box and I go on shaking, clearly, those cubes with the single different pole pointing to the outside will tend, with overwhelming probability, to mate with members of the other family, until my cubes have almost all paired up. Since the conditional probabilities of finding a member of family II, given the locus of a member of family I, has very much increased, the entropy of the system has gone down, hence we have more order after the shaking than before. It is easy to show* that in this case

* See Appendix.

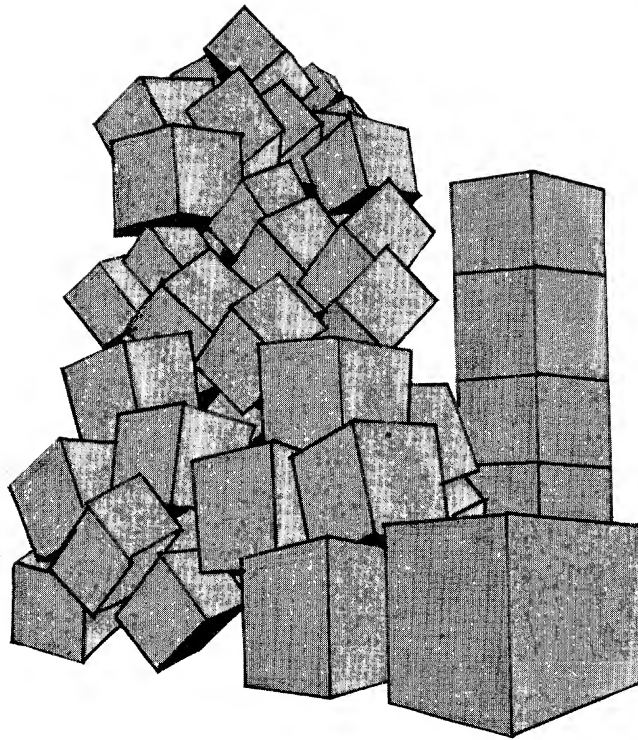


FIG. 5. Before.

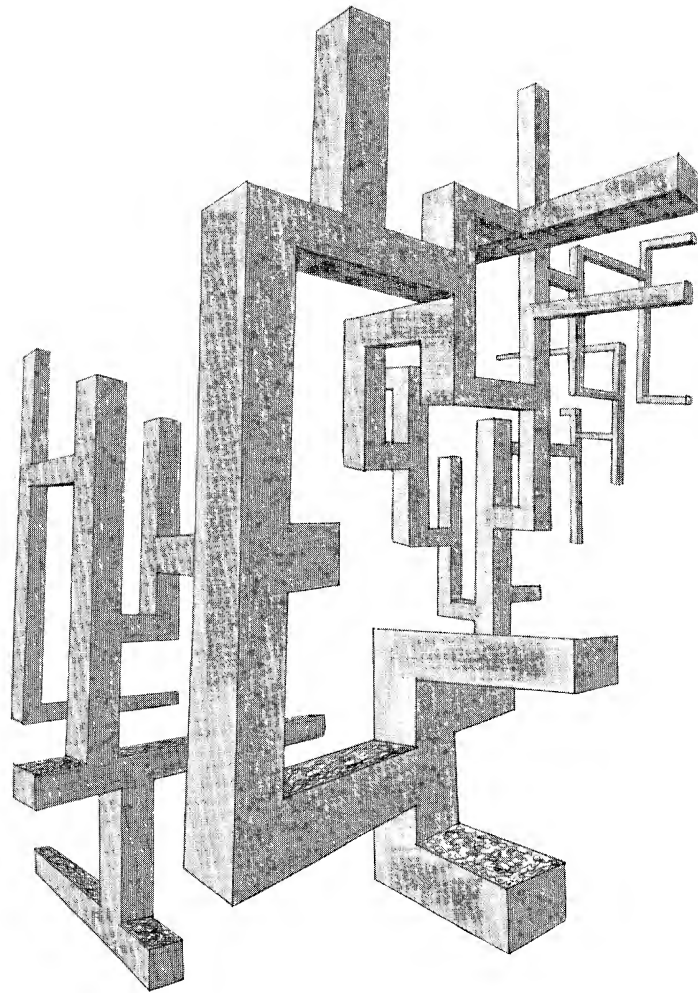


FIG. 6. After.

the amount of order in our system went up from zero to

$$R_{\infty} = \frac{1}{\log_2(en)},$$

if one started out with a population density of n cubes per unit volume.

I grant you, that this increase in orderliness is not impressive at all, particularly if the population density is high. All right then, let's take a population made up entirely of members belonging to family IVB, which is characterized by opposite polarity of the two pairs of those three sides which join in two opposite corners. I put these cubes into my box and you shake it. After some time we open the box and, instead of seeing a heap of cubes piled up somewhere in the box (Fig. 5), you may not believe your eyes, but an incredibly ordered structure will emerge, which, I fancy, may pass the grade to be displayed in an exhibition of surrealist art (Fig. 6).

If I would have left you ignorant with respect to my magnetic-surface trick and you would ask me, what is it that put these cubes into this remarkable order, I would keep a straight face and would answer: The shaking, of course—and some little demons in the box.

With this example, I hope, I have sufficiently illustrated the principle I called "order from noise," because no order was fed to the system, just cheap undirected energy; however, thanks to the little demons in the box, in the long run only those components of the noise were selected which contributed to the increase of order in the system. The occurrence of a mutation e.g. would be a pertinent analogy in the case of gametes being the systems of consideration.

Hence, I would name two mechanisms as important clues to the understanding of self-organizing systems, one we may call the "order from order" principle as Schrodinger suggested, and the other one the "order from noise" principle, both of which require the co-operation of our demons who are created along with the elements of our system, being manifest in some of the intrinsic structural properties of these elements.

I may be accused of having presented an almost trivial case in the attempt to develop my order from noise principle. I agree. However, I am convinced that I would maintain a much stronger position, if I would not have given away my neat little trick with

the magnetized surfaces. Thus, I am very grateful to the sponsors of this conference that they invited Dr. Auerbach⁽⁶⁾ who later in this meeting will tell us about his beautiful experiments *in vitro* of the reorganization of cells into predetermined organs after the cells have been completely separated and mixed. If Dr. Auerbach happens to know the trick by which this is accomplished, I hope he does not give it away. Because, if he would remain silent, I could recover my thesis that without having some knowledge of the mechanisms involved, my example was not too trivial after all, and self-organizing systems still remain miraculous things.

APPENDIX

The entropy of a system of given size consisting of N indistinguishable elements will be computed taking only the spatial distribution

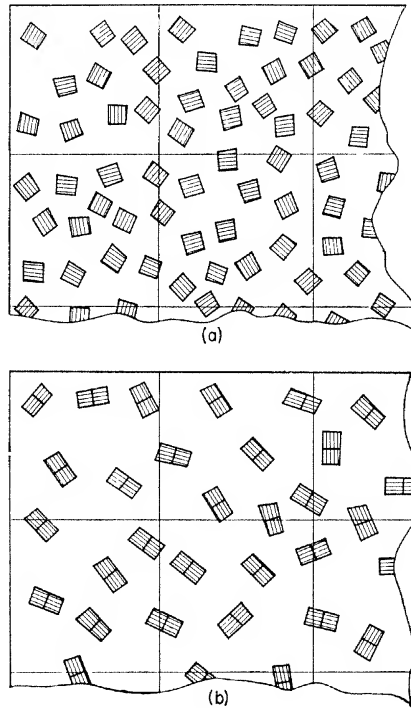


FIG. 7.

of elements into consideration. We start by subdividing the space into Z cells of equal size and count the number of cells Z_i lodging i elements (see Fig. 7a). Clearly we have

$$\sum Z_i = Z \quad (i)$$

$$\sum i Z_i = N \quad (ii)$$

The number of distinguishable variations of having a different number of elements in the cells is

$$P = \frac{Z!}{\prod Z_i!} \quad (iii)$$

whence we obtain the entropy of the system for a large number of cells and elements:

$$H = \ln P = Z \ln Z - \sum Z_i \ln Z_i \quad (iv)$$

In the case of maximum entropy $\bar{H}$ we must have

$$\delta H = 0 \quad (v)$$

observing also the conditions expressed in eqs. (i) and (ii). Applying the method of the Lagrange multipliers we have from (iv) and (v) with (i) and (ii):

$$\begin{array}{l|l} \sum (\ln Z_i + 1) \delta Z_i = 0 & \beta \\ \sum i \delta Z_i = 0 & \\ \sum \delta Z_i = 0 & - (1 + \ln \alpha) \end{array}$$

multiplying with the factors indicated and summing up the three equations we note that this sum vanishes if each term vanishes identically. Hence:

$$\ln Z_i + 1 + i\beta - 1 - \ln \alpha = 0 \quad (vi)$$

whence we obtain that distribution which maximizes H :

$$Z_i = \alpha e^{-i\beta} \quad (vii)$$

The two undetermined multipliers α and β can be evaluated from eqs. (i) and (ii):

$$\alpha \sum e^{-i\beta} = Z \quad (viii)$$

$$\alpha \sum i e^{-i\beta} = N \quad (ix)$$

Remembering that

$$-\frac{\delta}{\delta\beta} \Sigma e^{-i\beta} = \Sigma i e^{-i\beta}$$

we obtain from (viii) and (ix) after some manipulation:

$$\alpha = Z(1 - e^{-1/n}) \approx \frac{Z}{n} \quad (\text{x})$$

$$\beta = \ln \left(1 + \frac{1}{n} \right) \approx \frac{1}{n} \quad (\text{xi})$$

where n , the mean cell population or density N/Z is assumed to be large in order to obtain the simple approximations. In other words, cells are assumed to be large enough to lodge plenty of elements.

Having determined the multipliers α and β , we have arrived at the most probable distribution which, after eq. (vii) now reads:

$$Z_i = \frac{Z}{n} e^{-i/n} \quad (\text{xii})$$

From eq. (iv) we obtain at once the maximum entropy:

$$\bar{H} = Z \ln(en). \quad (\text{xiii})$$

Clearly, if the elements are supposed to be able to fuse into pairs (Fig. 7b), we have

$$\bar{H}' = Z \ln(en/2). \quad (\text{xiv})$$

Equating $\bar{H}$ with H_m and $\bar{H}'$ with H , we have for the amount of order after fusion:

$$R = 1 - \frac{Z \ln(en)}{Z \ln(en/2)} = \frac{1}{\log_2(en)} \quad (\text{xv})$$

REFERENCES

1. L. WITTGENSTEIN, *Tractatus Logico-Philosophicus*, paragraph 6.31, Humanities Publishing House, New York (1956).
2. G. A. PASK, The natural history of networks. This volume, p. 232.
3. C. SHANNON and W. WEAVER, *The Mathematical Theory of Communication*, p. 25, University of Illinois Press, Urbana, Illinois (1949).
4. W. HEISENBERG, *Z. Phys.* **43**, 172 (1927).
5. E. SHRODINGER, *What is Life?* pp. 72, 80, 80, 82, Macmillan, New York (1947).
6. R. AUERBACH, Organization and reorganization of embryonic cells. This volume, p. 101.

DISCUSSION

LEDERMAN (*University of Chicago*): I wonder if it is true that in your definition of order you are really aiming at conditional probabilities rather than just an order in a given system, because for a given number of elements in your system, under your definition of order, the order would be higher in a system in which the information content was actually smaller than for other systems.

VON FOERSTER: Perfectly right. What I tried to do here in setting a measure of order, was by suggesting redundancy as a measure. It is easy to handle. From this I can derive two statements with respect to $H_{\max}$ and with respect to H . Of course, I don't mean this is a universal interpretation of order in general. It is only a suggestion which may be useful or may not be useful.

LEDERMAN: I think it is a good suggestion but it is an especially good suggestion if you think of it in terms of some sort of conditional probability. It would be more meaningful if you think of the conditional probabilities as changing so that one of the elements is singled out for a given environmental state as a high probability.

VON FOERSTER: Yes, if you change H , there are several ways one can do it. One can change the conditional probability. One can change also the probability distribution which is perhaps easier. That is perfectly correct.

Now the question is, of course, in which way can this be achieved? It can be achieved, I think, if there is some internal structure of those entities which are to be organized.

LEDERMAN: I believe you can achieve that result from your original mathematical statement of the problem in terms of H and $H_{\max}$, in the sense that you can increase the order of your system by decreasing the noise in the system which increases $H_{\max}$.

VON FOERSTER: That is right. But there is the possibility that we will not be able to go beyond a certain level. On the other hand, I think it is favorable to have some noise in the system. If a system is going to freeze into a particular state, it is inadaptible and this final state may be altogether wrong. It will be incapable of adjusting itself to something that is a more appropriate situation.

LEDERMAN: That is right, but I think the parallelism between your mathematical approach and the model you gave in terms of the magnets organizing themselves, that in the mathematical approach you can increase the information content of the system by decreasing the noise and similarly in your system where you saw the magnets organizing themselves into some sort of structure you were also decreasing the noise in the system before you reached the point where you could say ah ha, there is order in that system.

VON FOERSTER: Yes, that is right.

MAYO (*Loyola University*): How can noise contribute to human learning? Isn't noise equivalent to nonsense?

VON FOERSTER: Oh, absolutely, yes. (Laughter). Well, the distinction between noise and nonsense, of course, is a very interesting one. It is referring usually to a reference frame. I believe that, for instance, if you would like to teach a dog, it would be advisable not only to do one and the same thing over and over again. I think what should be done in teaching or training, say, an animal, is to allow the system to remain adaptable, to ingrain the information in a way where the system has to test in every particular situation a hypothesis whether it is working or not. This can only be obtained if the nature into which the system is immersed is not absolutely deterministic but has some fluctuations. These fluctuations can be interpreted in many different forms. They can be

interpreted as noise, as nonsense, as things depending upon the particular frame of reference we talk about.

For instance, when I am teaching a class, and I want to have something remembered by the students particularly well, I usually come up with an error and they point out, "You made an error, sir." I say, "Oh yes, I made an error," but they remember this much better than if I would not have made an error. And that is why I am convinced that an environment with a reasonable amount of noise may not be too bad if you would really like to achieve learning.

REID (*Montreal Neurological Institute*): I would like to hear Dr. von Foerster's comment on the thermodynamics of self-organizing systems.

VON FOERSTER: You didn't say open or closed systems. This is an extremely important question and a very interesting one and probably there should be a two-year course on the thermodynamics of self-organizing systems. I think Prigogin and others have approached the open system problem. I myself am very interested in many different angles of the thermodynamics of self-organizing systems because it is a completely new field.

If your system contains only a thousand, ten thousand or a hundred thousand particles, one runs into difficulties with the definition of temperature. For instance, in a chromosome or a gene, you may have a complex molecule involving about 10^6 particles. Now, how valid is the thermodynamics of 10^6 particles or the theory which was originally developed for 10^{23} particles? If this reduction of about 10^{17} is valid in the sense that you can still talk about "temperature" there is one way you may talk about it. There is, of course, the approach to which you may switch, and that is information theory. However, there is one problem left and that is, you don't have a Boltzmann's constant in information theory and that is, alas, a major trouble.

STATISTICAL MODELS FOR RECALL AND RECOGNITION OF STIMULUS PATTERNS BY HUMAN OBSERVERS

W. K. ESTES

*Indiana University**

THE particular self-organizing system that is to be considered in this study is a human observer engaged in learning differential responses to stimulus patterns. These differential responses, providing the means of testing for recall and recognition, are a form of behavior that is a focus of interest for a number of disciplines. Consequently, it is not surprising to find a symposium such as the present one including several very different theoretical approaches to the same cluster of problems. Most of these approaches, perhaps all except the one I represent, take it for granted that we know well enough what is meant by "learning to differentiate stimulus patterns" and proceed to seek explanations, usually in terms of physical or neural models. Our one deviant, the experimental psychologist, instead makes a particular point of actively and persistently questioning whether we do indeed have a clear idea of just what is learned when an individual acquires a differential response to a stimulus pattern.

One might think it a simple matter to bypass niceties of definition by adopting the role of a hard-boiled operationalist. Why should we not simply agree to define the learning of a stimulus pattern in terms of the individual's ability to give the correct name when the pattern appears in a recognition test? The difficulty is that, as always when one puts on blinkers, we would be in danger of failing

* This paper was prepared during the writer's tenure as Visiting Professor of Psychology at Northwestern University, spring quarter 1958-59.

entirely to see important aspects of the problem. As Campbell⁽¹⁾ has brought out clearly in another context, what we choose to regard as a simple instance of learning involving but the strength of a single response may nevertheless represent much more than this from the viewpoint of the individual who is doing the learning. More concretely, other behavioral dispositions than the one we have decided to observe may change systematically during the learning of "a single response." The individual who has learned to associate a name with a stimulus pattern may differ from one who has not learned, in his ability to give the correct name when the pattern recurs, and also in his tendency to give this name when some portion of the original pattern appears in a new situation. An adequate theory of pattern learning and perception must not only describe the course of acquisition, but it must also yield predictions as to the likelihood that a learned pattern will be recognized under changed circumstances.

The research program to be discussed in the remainder of this paper represents a combined experimental and theoretical approach to the development of a quantitative account of pattern learning and recognition. In the experimental work, the dependent variables are empirical probabilities of response to stimulus patterns and their constituents. The theoretical strategy is to start by applying concepts and assumptions drawn from contemporary statistical learning theory,^(2,3,4) then to modify these in the light of feedback from the associated experimental analyses.

Two theoretical notions we shall require are that of a *stimulus element*, or cue, and that of *reinforcement*. A stimulus element is to be the conceptual counterpart of any portion of the stimulating environment which may enter into a unitary relation with a response during learning. The type of learning situation we shall be concerned with is that known to the psychologist as paired-associates. The training procedure is of the type used, for example, in aircraft recognition training. On each training trial, a stimulus pattern is presented to the learner together with an indication of the correct name; this combined occurrence of stimulation and correct response we denote as reinforcement. On a test trial, the learner is confronted with the stimulus pattern presented alone, or some portion of it, and attempts to give the correct response.

In order to express our assumption about the manner in which the probability of a response to a stimulus element changes as a function of reinforcement, we introduce the following notation. Suppose that a learning situation involves a collection of N stimulus elements ($e_1, e_2, \dots, e_N$) and a set of r alternative responses ($A_1, A_2, \dots, A_r$). By $p_{hi,n}$ we denote a theoretical quantity which may be regarded as the probability that stimulus element e_h , if presented alone on trial n of the experiment, would evoke response A_i . When we are dealing with response probabilities at the asymptote of learning we shall drop the subscript n . The law of reinforcement we shall take over from statistical learning theory states that if on trial n of an experiment, stimulus element e_h was present and sampled ("perceived") by the learner, and response A_i was reinforced, then

$$p_{hi,n+1} = p_{hi,n} + c(1 - p_{hi,n}), \quad (1)$$

where c is a constant with a value between zero and one. Equation (1) states that on each training trial the probability of the reinforced response to the given stimulus increases by a constant fraction of the difference between its current value and unity. This difference equation may be rewritten in the form

$$p_{hi,n+1} = (1 - c)p_{hi,n} + c \quad (1a)$$

if we wish to emphasize the linearity of the transformation.

Numerous experimental studies (e.g. Ref. 3) have shown that this function describes the course of acquisition in various simple learning situations in which the entire stimulating situation is replicated from trial to trial. Consequently we should expect it to hold also for the simple paired-associate problem, comprising a set of N distinct stimulus patterns and r distinct responses, as illustrated in the paradigm below.

<i>Stimulus</i>	<i>Response</i>
e_1	A_1
e_2	A_2
.	.
.	.
.	.
e_N	A_N

A cycle of training trials consists of a single reinforcement of the correct response to each of the N stimuli, the list being given in a new random order on each cycle. On the assumption that eq. (1) describes the change in probability of a correct response to each item on any one trial, it is easy to show by induction that probability of a correct response to any item can be expressed as a function of trials by the formula

$$p_{hi,n} = 1 - (1 - p_{hi,1})(1 - c)^{n-1} \quad (2)$$

If the value of the parameter c is the same for all items, then eq. (2), with the subscript h dropped, should also describe the expected proportion of correct responses per trial. A straightforward probabilistic derivation⁽⁴⁾ shows further that the expected proportion of instances in which k correct responses are given on the n th presentation of the list is expressed by

$$P_n(k) = \binom{N}{k} [1 - (1 - p_1)(1 - c)^{n-1}]^k (1 - p_1)^{N-k} (1 - c)^{(n-1)(N-k)} \quad (3)$$

Applications of eq. (2) and (3) to data from simple paired-associate experiments^(3,5) lend empirical support to our assumption that the course of simple associative learning is accurately described by eq. (1) and its various corollaries. In the remainder of this paper we shall be interested in this mathematical model, not for its own sake, but for its value as a tool of analysis.

In the simple paired-associate experiments, a given stimulus pattern was always presented as a whole, both on training trials and on recall tests. When we consider possible variations in this situation, one of the first questions that comes to mind is whether the effect of the training is to establish an association between a response and a stimulus pattern as a whole or also to establish associations between the response and various components into which the pattern can be analyzed. Consider, for example, the following item which occurred in a recent paired-associate experiment conducted by Judith P. Frankmann⁽⁶⁾ in our laboratory.

<i>Exper. A</i>	<i>Stimulus</i>	<i>Response</i>
	<i>rc</i>	1
	<i>gi</i>	2

Stimulus pattern *rc* consisted of a red light presented together with a continuous tone and response 1 was always reinforced in its presence. Stimulus pattern *gi*, consisting of a green light together with an intermittent tone, was similarly correlated with reinforcement of response 2. By the end of a training series comprising 30 reinforced trials on each pattern, a group of 24 human subjects was making nearly 100% correct responses. Evidently each of the two patterns could be regarded as a stimulus element for which eq. (1) was satisfied. We may ask, however, whether the same was true also of the separate components, *r*, *g*, *c*, and *i*.

To gain evidence on this point, tests were given with these components appearing alone. The results of these tests are shown in Table 1, the "correct" response to a component being designated as the one previously reinforced to the pattern of which it was a constituent.

TABLE 1. TEST DATA FROM FRANKMANN STUDY

Stimulus	Proportion of "correct" responses
<i>rc</i>	0.94
<i>r</i>	0.83
<i>c</i>	0.68
<i>gi</i>	0.81
<i>g</i>	0.78
<i>i</i>	0.62

(It should be mentioned that the symbols given here should not be taken literally, for the data are pooled over a number of different sets of tones and lights, all of which had been manipulated analogously to those shown in Table 1.) On the whole, the proportions of correct responses to components are somewhat below those to the corresponding patterns.

The appropriate interpretation is not immediately obvious. One's first reaction might be to conclude that the association of a response with a complete training pattern is stronger than its association with the separate components. However, we should note that the tests with components alone must have involved some new stimulation that had not been present during training and that would be expected to act as "noise" in producing regression of the

observed response proportions toward a chance level. With this consideration in mind, we should perhaps suspend judgment for the moment and seek additional tests of the hypothesis that the responses have become associated with the stimulus components in the same way and to the same degree as with the full patterns.

In order to obtain some independent evidence on this point, let us turn to a new, and slightly more complex experimental situation. In the Frankmann experiment, both a given pattern and its constituents were always associated with the same response. This special restriction is not an essential aspect of paired-associate training, however. More generally, we expect human subjects to achieve accurate performance in the recognition of stimulus patterns even if the patterns are composed of components which vary independently.

Consider for example the situation used by Arnold Binder for a series of studies of recognition training in the Indiana laboratory. The experimental paradigm took the following form:

<i>Exper. B</i>	<i>Stimulus pattern</i>	<i>Response</i>
	<i>a c e</i>	A_1
	<i>a d e</i>	A_2
	<i>a c f</i>	B_1
	<i>a d f</i>	B_2
	<i>b c e</i>	C_1
	<i>b d e</i>	C_2
	<i>b c f</i>	C_3
	<i>b d f</i>	C_4

The small letters in the left-hand column represent components of the stimulus patterns, which were actually geometrical figures; *e* and *f* denote a triangle and a circle, respectively, *a* and *b* a horizontal straight or jagged line above the triangle or circle, *c* and *d* an oblique straight line at the lower left or right corner of the figure (for additional details, see Ref. 7).

During training, each of the eight patterns is uniformly associated with reinforcement of the response paired with it. With sufficient practice, human subjects attain 100% accuracy in recognizing these patterns and give the correct responses when the patterns are presented on test trials. Suppose, however, that an incomplete pattern, say *a e*, or *b* alone, is presented on a test trial. If a single

component, or a sub-pattern, acts as a unit ("stimulus element") in this situation, then the theoretical assumption derived from the simpler experiment leads us to predict that the subject's response to b alone will be either C_1 , C_2 , C_3 , or C_4 , and his response to a will be either A_1 or A_2 . The results of a number of tests of this kind⁽⁷⁾ show that predicted values are approached although there is a small attenuation in the probabilities of "correct" responses when tests are given in new stimulus situations that had not occurred during training.

Our analysis of this last situation is not yet complete, for one may ask, not only which responses will be given to a sub-pattern on a recognition test, but also how response probability will be divided among the "appropriate" alternatives. To deal with this question, we require two theorems derivable from the law of reinforcement (eq. 1) for less restricted boundary conditions than those involved in our treatment of the simple paired-associate problem. We consider now a class of situations, of which Binder's is a special case. One response from a set of r alternatives is reinforced on each trial, and the relative frequencies of stimuli and reinforcing events are so arranged that when stimulus element e_h occurs on a training trial, response i is reinforced with probability π_{hi} (π_{hi} is constant over the training series). By a derivation that has been published elsewhere,^(2,4) our law of reinforcement yields the following difference equation

$$p_{hi,n+1} = (1 - c)p_{hi,n} + c\pi_{hi} \quad (4)$$

as the function describing the expected change in p_{hi} from trial to trial under the specified conditions (a trial now being interpreted as any training trial on which stimulus element e_h is present). We shall be interested primarily in the asymptotic value p_{hi} of the conditional response probability. By setting $p_{hi,n+1} = p_{hi,n} = p_{hi}$ in eq. (4) and solving for p_{hi} this value is found to be

$$p_{hi} = \pi_{hi}. \quad (5)$$

Asymptotically, the conditional probability of a given response to a stimulus element approaches the conditional probability of its reinforcement in the presence of that element during training. This result we shall refer to as the *matching theorem*.

Now we are ready to return to the experiment on pattern recognition. In the particular study discussed above (Exper. B), all eight of the patterns were presented equally often during training. Therefore, we should predict that, for example, A_1 and A_2 would occur equally often in response to the test combination $a e$, and that C_1 , C_2 , C_3 , and C_4 would occur with equal frequencies in response to the test on b alone. These expectations were borne out by the data.

A more interesting application of the matching theorem arises if the various patterns are presented with unequal frequencies during training. Suppose, for example, that the pattern $a c e$ occurred twice as often as $a d e$. Then, if the matching theorem is satisfied with the sub-pattern $a e$ serving as a stimulus element, the conditional probabilities of reinforcement of responses A_1 and A_2 in the presence of this element would be $\frac{2}{3}$ and $\frac{1}{3}$, respectively. Asymptotically, these values should be matched by the conditional response probabilities. For a group of 55 subjects in one of Binder's experiments, these relative frequencies did in fact obtain. The results of the test on $a e$ are shown below:

<i>Response</i>	<i>Observed proportion</i>	<i>Theoretical proportion</i>
A_1	0.65	0.67
A_2	0.29	0.33
Other	0.06	—

In the same study, the third and fourth patterns, $a c f$ and $a d f$, appeared with relative frequencies in the ratio 4:1 during training. On a recognition test with sub-pattern $a f$, the response proportions were:

<i>Response</i>	<i>Observed proportion</i>	<i>Theoretical proportion</i>
B_1	0.76	0.80
B_2	0.20	0.20
Other	0.04	—

These and other similar results from an extensive series of studies by Binder and his associates,⁽⁷⁾ as well as related experiments conducted by Peterson,⁽⁸⁾ Solley and Messick,⁽⁹⁾ and by the writer, have provided an impressive accumulation of evidence suggesting that eq. (1) and the matching theorem are satisfied for the whole hierarchy of sub-patterns and components which vary

as units during a series of learning trials; the probability of a response to any sub-pattern when tested alone tends asymptotically to the marginal relative frequency with which the response has been reinforced in the presence of the given sub-pattern during learning.

We must hasten to add, however, that this elegantly simple and clean-cut conclusion is justified only when training is conducted under the type of stimulus arrangements we have considered. To bring out an important qualification and at the same time to leave the reader with a realistic picture of the present state of research in this area, I cite one more empirical finding. Below we see the experimental paradigm for a study of human discrimination learning recently conducted in my laboratory with the assistance of B. L. Hopkins.

<i>Exper. C</i>	<i>Relative frequency</i>	<i>Stimulus</i>	<i>Response</i>
	1/8	<i>a</i>	1
	1/8	<i>b</i>	2
	1/4	<i>a b</i>	1
	1/8	<i>c</i>	1
	1/8	<i>d</i>	2
	1/4	<i>c d</i>	2

The first column gives the relative frequency with which the stimulus in the second column occurred during training, and the third column gives the response reinforced in the presence of the given stimulus. The component stimuli were actually Russian script characters, but for convenience we represent them here by letters available on an American typewriter. The responses consisted of an upward and a downward movement of a lever.

Training continued for 128 trials, at the end of which the curve of correct response proportions was clearly approaching unity for each of the six training stimuli. This result, taken by itself, is not surprising, but it does raise a difficulty for the conclusion about probability matching drawn from the preceding experiments. The asymptotic probability of response 2 to stimulus *b*, for example, approaches the relative frequency of reinforcement in the presence of *b* alone, not the marginal relative frequency with which response 2 was reinforced whenever *b* was present (i.e. on *b* and *a b* trials taken together). A similar statement holds for the probability of response 1 to stimulus *c*.

The substance of these results would seem to be that when a cue appears both as a separate pattern and also as a component of a larger pattern, with different reinforcement contingencies on the two types of trials, the probability of any given response to this cue satisfies eq. (1) and the matching theorem over the sub-sequence of trials on which the cue appears as a pattern alone. Put differently, once a response has become associated with a stimulus pattern appearing as a whole, this association is undisturbed by reinforcing events that occur on trials when this stimulus appears as a component of a larger pattern. Through the same operation of reinforcement, a response may become associated with a pattern as a unit, with the constituents of a pattern, or under some circumstances, with both simultaneously. However, in a sense, the relation involving the entire pattern as a unit is dominant over those involving its components.

Considering the present state of research on pattern learning, it seems more fitting for a progress report to terminate with a problem than with a conclusion. To this end, let us refer once more to the paradigm of the last experiment (Exper. C). Suppose that after the termination of discrimination training, we assemble some of the component stimuli into new test patterns. We might, for example, present abc as a test pattern. What should we predict for the probability of response 1? If the sub-pattern ab acts as a unit, then evidently the probability of response 1 should be unity, since it is the response reinforced both to ab and to c during training. If, however, a and b act as units separately, then the probability of response 1 should be less than unity, perhaps $\frac{2}{3}$ since two of the three elements have become cues for response 1 during training. Similar considerations would obtain for a number of other test patterns, abd and acd , etc.

These tests have been actually conducted and the results will be reported in due course along with my own theoretical analysis of the test situation. In the meantime, other investigators may wish to avail themselves of the opportunity of testing their theories of pattern learning and recognition by formulating predictions of the response proportions on these tests without any danger of bias from prior knowledge of the results. Leaving this problem hanging in air will also serve to underscore the fact that even in relatively simple experimental situations one may still encounter areas of real

uncertainty as to precisely "what is learned" when an individual acquires differential responses to stimulus patterns.

REFERENCES

1. D. T. CAMPBELL, Operational delimitation of "what is learned" via the transposition experiment. *Psychol. Rev.* **61**, 167-74 (1954).
2. C. J. BURKE and W. K. ESTES, A component model for stimulus variables in discrimination learning. *Psychometrika* **22**, 133-46 (1957).
3. W. K. ESTES, The statistical approach to learning theory. In *Psychology: A Study of a Science*. Vol. 2 (ed. by S. KOCH). McGraw-Hill, New York (1959).
4. W. K. ESTES, Component and pattern models with Markovian interpretations. In *Studies in Mathematical Learning Theory* (ed. by R. R. BUSH and W. K. ESTES). Stanford University Press, Stanford, California (1959).
5. W. K. ESTES, The growth and function of mathematical models for learning. In *Current Trends in Psychology* (ed. by R. GLASER). University of Pittsburgh Press, Pittsburgh, to appear in 1960.
6. J. P. FRANKMANN, Discrimination learning with single, compound, and alternate stimulus sets. Ph.D. dissertation, Indiana University (1959).
7. A. M. BINDER and S. FELDMAN, *Learning Factors in Recognition*, *Psychol. Monogr.*, in press.
8. L. R. PETERSON, Prediction of response in verbal habit hierarchies. *J. Exp. Psychology* **51**, 249-52 (1956).
9. C. M. SOLLEY and S. J. MESSICK, Probability-learning, the statistical structure of concepts, and the measurement of meaning. *Amer. J. Psychology* **70**, 161-73 (1957).

DISCUSSION

NEWELL (*Rand Corporation*): I would like you to comment on whether in this last situation this is a learning situation for the subject or whether this is a dilemma resolution situation for the subject. It seems to me that in some sense he is faced with an entirely new stimulus like an ABC sign which he has never seen before and his problem is what to do. He is forced to respond. He doesn't see anything in the pattern that he has known in the past and so he has to select one thing. Maybe one of the major issues is what the experimenter wants him to do in the situation, and you would not really characterize this as a learning situation proper.

ESTES: It is hard to know in which of many possible senses to answer the question. I prefer to regard it as a learning situation for several reasons. One is that I happen to be identified with the psychology of learning personally. A second one is that I don't see any immediate likelihood that it would be profitable to apply other kinds of theories such as decision theory to the subject's task in the situation because he receives this test combination once at the end of training and he has no opportunity to receive feedback from his behavior on these test trials and modify it in the light of experience or in the light of principles or what not.

He is given a single test on a new pattern with a very short time to respond and it is hard to see how his response could be affected by anything but the immediately preceding pattern which would, for me, by definition, make this a problem for learning theory rather than a problem for decision theory.

MINSKY (*M.I.T.*): Could you say a word or two about this matching theorem which is a lovely experimental result? It falls right out of the theory. But it seems to me like a rather inefficient strategy for any really adaptive learning creature to use. That is, if you get rewarded 90% of the time for something you ought to apply it all the time. I wonder how you could fit that into a more complicated theory of behavior?

ESTES: I could make a couple of comments on that. Firstly, in an extremely simple situation where there are say only two alternatives being presented repeatedly and having different probabilities of success, then it is indeed an inefficient strategy for the learner to follow the matching theorem rather than to follow the statistical decision theory, or a result that would be predicted from game theory. It is interesting to note, however, that it is precisely and almost only in these very simple choice situations that this matching theorem appears to break down after a long series of trials. In some simple two choice situations, after a long series of trials, apparently something else comes in and the subjects depart from the matching theorem and do go to 100%.

As soon as we put them into a more difficult situation such as this recognition problem with all its independently varying cues, then, in my experience, the matching theorem is really accurately satisfied numerically. It may be that this is better than anything else that could easily be offered.

MAYO (*Loyola University*): In your first example when you presented rc as a pair, then r or c alone, how much savings was there for either r or c ? I mean savings in the Ebbinghaus sense.

ESTES: I presume this would mean savings in the trials required to learn to criteria on the new stimuli. That I can't answer because we gave no additional training on the component stimuli in our experiment. More generally, what one would want to know would be how much retention decrement there was on the components alone and that in absolute terms wouldn't mean anything, so the problem comes down to expressing a decrement as a function of the conditions that are responsible for it. Maybe if we have another conference in ten years we can add something useful to that.

LYMAN (*UCLA*): Does the matching theorem work out when the responses have a differential reward value? For example, one response gets a very large reward and the other a very small reward?

ESTES: The briefest answer is no. What I think is the case, at least for a fair class of experiments in human subjects, is the following: If the consequences of two choices have different rewards, for example different amounts of money or something of the sort, over a series of trials one does not satisfy the matching theorem as stated, but possibly one can find a pair of numerical weights to be associated with these two differential rewards such that they will then satisfy a modified matching theorem for any probabilistic schedule you might wish to put them in. If so, that would mean one could use learning experiments of this sort for the purpose of measuring utilities, as the economists try to do from a different viewpoint.

PERCEPTUAL GENERALIZATION OVER TRANSFORMATION GROUPS*

FRANK ROSENBLATT

Cornell Aeronautical Laboratory, Inc., Buffalo, New York

I. INTRODUCTION

THIS paper is concerned with the question of how a brain, or brain-like system, can recognize *similarity* among the various possible transformations of a sensory pattern, or image. If we make an arbitrary choice of a transformation group—say, the group of all rigid motions—then we are interested in recognizing as *similar*, and applying the same name, or eliciting the same response to, all patterns which are generated by applying the transformations of the group to some arbitrary initial configuration. To avoid triviality, we require at the same time that the system assign *different* names or responses to stimuli which are *not* similar under the transformation group in question. For convenience of discussion, we shall limit our remarks to those transformations which can be represented as one-to-one mappings of a visual field, or retina, on to itself. This does not mean that the methods discussed are necessarily limited either to vision or to one-to-one transformations; the constraint is introduced only to enable us to deal more rigorously with a well-defined class of problems. After reviewing three fundamental methods hitherto suggested for dealing with this problem, I would like to take this opportunity to introduce a fourth fundamentally different method, which has come to light only during the last few months, in the course of work on the perceptron program, at the Cornell Aeronautical Laboratory. (See Refs. 1, 2 and 3 for a general review of this program.) In presenting this fourth method, I hope to show how a self-organizing system can be designed which is capable of abstracting from a given environment those transformations which most frequently occur, and which will

* This work was supported jointly by the Rome Air Development Center and the Office of Naval Research under O.N.R. Contract Nonr. 2381(00).

then be capable of recognizing the similarity or dissimilarity of entirely new, hitherto unexperienced patterns, in terms of these transformations.

II. ADMISSIBLE TRANSFORMATIONS

Figure 1 illustrates a number of transformations of the letter "N" which would be admissible under the class of 1:1 transformations being considered here.* The various recognition methods which we shall consider vary considerably in their capability for dealing with these different transformations, as we shall see. It is of particular importance to refrain from thinking of a transformation as necessarily implying an operation *on the stimulus image*; in each case, the transformation is an abstract operator applicable to the *entire retinal space*, whether or not any particular pattern of illumination happens to exist. Thus, it is meaningful to talk about applying the *same* transformation (e.g. a rotation of 90°) to any one of a number of different stimuli (such as the letter "X" or the projection of North America). It should also be noted that several geometric problems are immediately introduced if we try to satisfy the requirements of a 1:1 transformation and at the same time to maintain a flat, bounded Euclidean space as the retinal field. Rigid motions, for example, would be ruled out in such a system, because points close to a boundary would be carried outside the retinal space. In order to allow for the possibility of rigid motions (which can be most readily simulated and analyzed), it is expedient in some of our models to assume a decidedly nonbiological retina, having a toroidal connectivity (or Born-von Kármán boundary conditions), so that a stimulus pattern which is shifted off one edge simply re-enters at the opposite side of the field. More serious problems are introduced if we try to represent the projections of a three-dimensional physical space upon a finite two-dimensional retina; for example, the masking of one form by another, and its subsequent re-emergence, cannot be represented in a system which is subject to the 1:1 constraint. Although we are currently investigating some of the formal problems in the representation and analysis of such events, we will not enter into this here.

* Actually, only those 1:1 transformations under which the measure of the set of fixed points is close to zero will be considered here. Rotations, or distortions about a single fixed point are considered admissible, but the identity transformation (for example) is not.

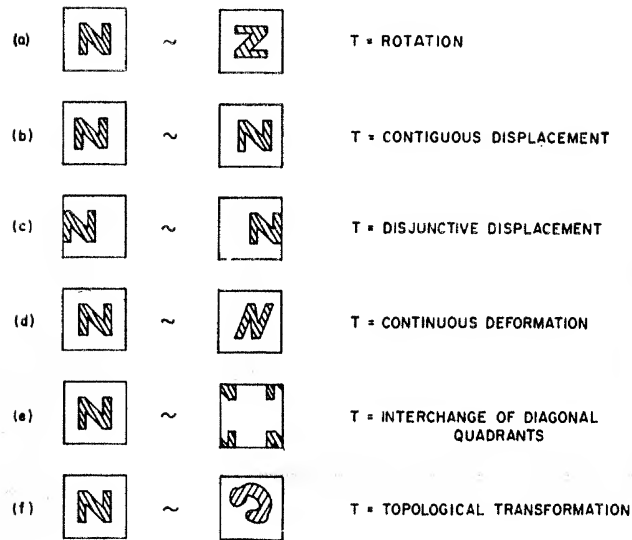


FIG. 1. Some possible 1:1 transformations.

III. METHODS PREVIOUSLY PROPOSED

Let me propose four names, for convenience of reference, for the four methods of transformation-recognition which are to be considered. These are:

1. The analytic-descriptive method.
2. Image transformation.
3. Contiguity (or preponderance) generalization.
4. Transform association.

The *analytic-descriptive method* consists of reducing a stimulus pattern, or configuration, to a simple, canonical description, which is invariant under the transformations in question. This description (generally given in terms of measurements of lines and angles, ratios of dimensions, etc.) can then be compared with a stored set of master-descriptions to determine which corresponds most closely to the stimulus on hand. (See particularly, Refs. 4 and 5 for typical applications of this approach.) In general, these systems are neither neurologically oriented nor are they self-organizing. The usual objective is to find some method of analysis by which figures can be described, and which is applicable to programmed application

either by a digital computer or by some special-purpose device. A possible exception to this statement is the system proposed by Oliver Selfridge⁽⁶⁾ where a given vocabulary of elementary analytic operations and functions are experimentally recombined and optimized through experience to yield a desired recognition function in a more or less arbitrary universe of forms. We might also mention Hebb's "phase sequences"⁽⁷⁾ as a neurological concept which probably best fits in this class. Most such systems, however, are clearly special-purpose devices, which are as lacking in simplicity of organization as they are in generality. Nonetheless, these are the only systems, to date, which can deal readily with the recognition of certain topological transforms (see Fig. 1f), and it seems quite plausible that a more advanced system than our simple perceptrons might begin to make use of such techniques.

The method of *image transformation* has been more seriously advanced as a possible model for the brain's operation in the perceptual-generalization problem, or, as McCulloch and Pitts call it, the problem of knowing "universals."^(8,9) In this method, a network of neural elements (or similar components) is devised, which actually applies all of the admissible transformations to a stimulus image which is obtained from the retina, and attempts to "normalize" this image in position, size, angular rotation, etc. so that at some point the image can be superimposed on, and recognized as identical to one of a number of stored normalized forms, or "memory traces." This method has a number of serious disadvantages. First of all, it is apt to be most uneconomical in the number of neural elements required. Some of Culbertson's models, for example,⁽⁸⁾ could readily use up all of the cells of the human central nervous system just in the process of centering, rotating, and transforming the size of a retinal image. Each new transformation which we wish to consider (such as elongation, or a tilting-distortion of the type that would produce italics from standard characters) requires an additional specially designed computing network, or transformation network. Such networks require a precision of connections and a type of functional specificity of different units which seems to be most uncharacteristic of biological systems; they do not seem to mesh at all well with the observed ability of an undamaged part of the cerebral cortex to take over the functions of damaged areas. Moreover, once a particular

transformation function is built into the system, it *must* be applied, willy-nilly, to whatever form comes in. Thus, in Fig. 1(a), a system which rotates its stimuli would automatically assign both patterns to the same class, and could never learn that an "N" rotated 90° is not to be called an "N" but a "Z". On the other hand, if the system does *not* have a built-in mechanism for rotation, it could never recognize the potential similarity of an "N" in normal position and one which is tilted, even slightly. These logical difficulties can, in most cases, be overcome with sufficient inventive ingenuity, but only at a frightful cost in simplicity, and I am convinced that, in a field which has yet to match even the most elemental products of genetics and evolution, a theoretical brain model should, if nothing else, obey the canon of simplicity.

Now what, exactly, does simplicity mean, in this context? Surely Warren McCulloch would claim that his neurons, which are simply one-shot triggered pulse generators, and which are the sole logical constituent of his nerve nets, are very simple devices, and I would be the first to agree with him. But if we consider the information which would have to be provided in a genetic system in order to grow a McCulloch-Pitts nerve net from scratch, it turns out that the specification requirements are likely to be quite formidable. This is a reflection of the fact that the degrees of freedom of a logically determined computing-network are apt to be quite limited, in the sense that any major random perturbation upon the predesigned plan would be likely to cause serious malfunctions in performance. When we use the term "simplicity" in talking about a nerve net, we do not, in general, refer to the number of elements, or to the number of variables in their individual functional equations, but rather to the degree of constraint upon the system as a whole. It was with the idea of studying the evolution of learning and memory in minimally constrained systems that the perceptron program was initiated.

It should be noted that for any given perceptron, regardless of which table of random numbers was used in its construction, it would be possible in principle to write a complete McCulloch-Pitts logical equation, describing the set of all possible states of the system as a logical function of the set of all possible inputs. We may ask, therefore, what has been gained by the use of statistical rules of organization rather than writing the logical equation in

the first place. The fact is, that the logical specification equation is a particularly idiosyncratic function of the specific network and is apt to be totally different for two networks having identical performance characteristics in the domain of interest. In fact, each perceptron which is constructed in accordance with our statistical rules is likely to yield a completely different logical equation; the states of the system produced by identical stimuli would almost certainly be strikingly different in every case. Nonetheless (if the systems are large enough) those functional characteristics which we are most concerned with, such as the ability correctly to discriminate a particular pair of forms would be found to vary little, so long as the statistical rules remain unchanged. It appears, therefore, that the statistical rules come closer to a canonical specification of what is most important for the systems to operate properly. The number of *logical* specifications which fit the bill is astronomical, and for the most part these constitute interchangeable variations on a theme; but any violation of the *statistical* structure of a perceptron is likely to radically alter the performance of the system.

This digression on simplicity and statistical organization now brings us back to the third method for transformation recognition, which I have termed *contiguity generalization*, and which Clark and Farley⁽¹⁰⁾ have called *preponderance generalization*. This is the method which has previously been employed in the perceptron. In its simplest form (as demonstrated by Clark and Farley) this method simply rests on the fact that a slight displacement or modification of a pattern, which shares most of its sensory points with the original pattern (as in Fig. 1b), is likely to evoke the same response as the original, rather than an alternative response which is associated with a different pattern with which it has few points in common. In our early work on the perceptron^(1,2) it was shown that the same response can be generalized over a class of forms many members of which are actually completely disjunct from one another on the retina. In order for this effect to work, however, it is necessary that each new stimulus pattern share a set of retinal points with at least one stimulus to which the response has previously been associated. Thus, in Fig. 1c, if we wish a response which has been assigned to the figure on the left to be generalized to the disjunct image on the right, it is necessary for the perceptron to see

the letter "N" in a large number of intermediate positions, so as to form a chain, or "contiguity sequence," connecting the original figure to the new one. If this is done, and if the parametric design of the system is right, then we can show that the perceptron will spontaneously tend to generalize the response originally assigned to the "N" on the left, so that it eventually occurs for the figure on the right.

This "spontaneous organization" effect⁽³⁾ was originally demonstrated in a digital simulation experiment using squares randomly placed in the left half of the visual field as one class of stimuli, and squares randomly placed in the right half of the field as the other class to be discriminated. Since no squares were allowed to appear in the middle of the field, we actually had two disjunct classes, within each of which intersections of stimuli were possible. This experiment, however, is in a sense a trivial one, for the specification of any illuminated retinal point would be sufficient to specify the class to which the stimulus belongs, and we are not actually obtaining *form* discrimination at all, but only discrimination of position. A more sophisticated experiment was subsequently performed, using horizontal and vertical bars as the two stimulus classes. The perceptron used in this experiment is illustrated schematically in Fig. 2, and the performance curves obtained are shown in Fig. 3. Since the perceptron was assumed to have an infinite number of association units, it was possible to calculate the response of the system exactly instead of simulating it, unit by unit, as has been done in some of our other experiments. The system consists of a 20 by 20 retinal mosaic, in which a toroidal connectivity is assumed, as described above. The sensory points are connected at random to association units, and these are all connected to a single binary response. An association unit (or *A*-unit) may deliver either a positive or negative output signal, depending upon its past history. If the total signal received by the response (or *R*-units) is positive, the response $R = 1$ occurs; if the signal is negative, the response $R = 0$ occurs. If $R = 1$, then a reinforcement operator, ρ , is applied, as shown by the feedback line in Fig. 2. This reinforcement operator induces a gain in the strength of the output signals (or in the weight of the connections from *A*-units to the *R*-unit) for all of the active *A*-units. This gain is retained permanently, unless specifically altered by subsequent events. At the same time, any units which

are currently *inactive* at the time the reinforcement is applied *lose* in strength (algebraically) so that collectively they just cancel out the gain of the active units, keeping the total weight, or value, over

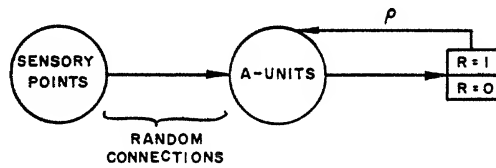


FIG. 2. A simple perceptron capable of contiguity generalization.

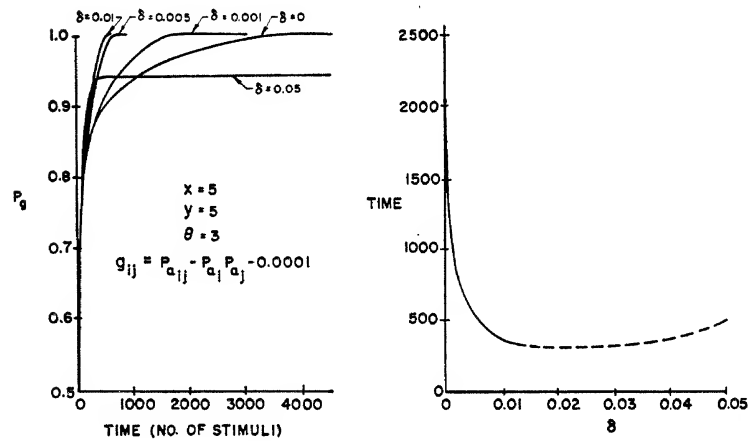


FIG. 3. Experiment 5-4.

the entire set of A -units equal to zero at all times. A system in which the gain in the active units is balanced by a compensating loss in the inactive units is called a *gamma system*. In the present system, a decay rate, δ , and a slight additional loss-component, ϵ , has been introduced, yielding the difference equation

$$\Delta v_i = \rho(\alpha_i^* - P_a - \delta v_i - \epsilon) \quad (1)$$

In this equation, v_i = the current output signal strength of the unit a_i . $\alpha_i^* = 1$ if a_i is active and 0 if a_i is inactive. $\rho = 1$ if $R = 1$ or 0 if $R = 0$. P_a is the proportion of active A -units.

The net effect of these dynamics is described by the "generalization coefficient,"

$$g_{ij} = P_{a_{ij}} - P_{a_i}P_{a_j} - \varepsilon \quad (2)$$

where $P_{a_{ij}}$ = the proportion of A -units responding both to stimulus S_i and to stimulus S_j

P_{a_i} = the proportion of A -units responding to S_i

and P_{a_j} = the proportion of A -units responding to S_j .

Equations for these functions have been given elsewhere.^(1,2) The generalization coefficient, g_{ij} , can be interpreted as a measure of the reinforcement which carries over to stimulus S_j as a result of presenting and reinforcing the stimulus S_i . A more complete discussion of this function will be found in the Appendix to Ref. 3, and in a forthcoming report. For the time being, it is sufficient to note that if g_{ij} is positive, then the stimulus S_j will tend to activate the same response which has been associated to S_i , while if g_{ij} is negative, S_j will tend to evoke the opposite response from that associated to S_i . It can be shown that two stimuli can be associated to the same response even if their g_{ij} is negative, but it is essential that the sum of the g_{ij} over all pairs of stimuli in a class must be positive, or else it is impossible to associate all members of that class to the same response.

The curves in Fig. 3 show the progress of learning for several different values of the decay rate, δ . δ is the rate at which previous reinforcements are lost, or "forgotten" by the A -units. The measure of performance used in this experiment is P_g , the probability of correct generalization. A P_g of 1 means that all of the horizontal bars evoke one response (either $R = 0$ or $R = 1$) while all of the vertical bars evoke the opposite response. P_g is equal to the proportion of the figures which are classified consistently. It should be emphasized that no attempt is made in the course of this experiment to direct the system or to influence it in any way in its choice of response; stimuli from the two classes are presented in a random sequence, and the only rule of operation is that whenever the response $R = 1$ occurs, the A -units are "reinforced" as described above. If the response is 0, no reinforcement is applied. Note that there is an optimum value of the decay rate, δ . If the decay is too small, the system becomes rigidly set in a "wrong" pattern of response, while if the decay rate is too great, the perceptron forgets too rapidly, and learning is unstable.

Now the results of experiments such as this were initially quite encouraging. Clearly, we had found a system which could spontaneously differentiate dissimilar forms, while assigning similar forms to the same class. It soon became evident, however, that a perceptron designed in this fashion is exceedingly temperamental, and will perform properly only under a limited spectrum of environmental conditions. In the experiment just illustrated, horizontal and vertical bars 4 units wide and 20 units long were used as stimuli. If, instead, we use a 4 by 20 vertical bar as one stimulus class, and a pair of parallel 2 by 20 bars, separated by a space of 3 units, and also vertical, as the other class, it can be shown that no combination of parameters in eq. (1) will enable a perceptron of the logical design just described spontaneously to form two different classes. It happens that in the conditions of the former experiment, the intersection of any pair of bars drawn from opposite classes is exactly equal to the expected value of this intersection over all possible pairs and, whenever this is true, the perceptron will act in a well-behaved fashion. If (as in the second case) we are dealing with two classes such that the intersection of stimuli from opposite classes can exceed its expected value, spontaneous discrimination will be difficult, and often impossible, even though under "forced learning"^(10,11) it may be quite easy to teach the perceptron to tell the stimuli apart.

Another peculiar problem of this type of system becomes evident from a consideration of the form of the learning curves, as illustrated in Fig. 3. These curves are convex, indicating increasing difficulty in completing the classes which have been correctly started, whereas the learning curves for a human subject, given the same problem, would certainly be concave, and in fact would rise directly to a maximum level of performance as soon as he "caught on" to the essential difference between the classes. The painfully slow rate of generalization which results from having to see stimuli in just the right position in order to generalize a step further from an existing "bridgehead" is also impressive. Such considerations as these strongly suggested that there must be some other method of generalization which could be employed by biological systems, without necessarily falling back upon the more complex and artificially constrained methods which were previously rejected.

Ultimately, this led to the discovery of the *transform association* method, which we will now proceed to develop more rigorously.

IV. TRANSFORM ASSOCIATION SYSTEMS ✓

In all of the perceptrons analyzed previously, we have assumed that all output connections from an *A*-unit must terminate on an *R*-unit. Basically, these perceptrons consist of a large number of parallel channels carrying signals from *S*-points to *R*-points. We will hereafter refer to such systems as *parallel connection perceptrons*. In contrast to such systems, we will now consider the class of *cross-coupled perceptrons*, in which an *A*-unit may have output connections terminating on other *A*-units as well as on *R*-units. The organization of such a system is illustrated schematically in Fig. 4. The rules of organization are as follows:

(1) Each *A*-unit is assumed to receive a fixed number, n_e , of excitatory connections, and a fixed number, n_i , of inhibitory connections from the sensory system. Each of these connections originates from some sensory point, chosen at random. The excitatory connections carry a unit positive signal, and the inhibitory connections carry a unit negative signal, if the sensory point from which they originate is illuminated. The set of sensory points connected to a given *A*-unit are called the sensory *origin points* of that *A*-unit. If no two connections to the same *A*-unit originate from the same point, then there are $n_e + n_i = m$ origins per *A*-unit. There are no constraints on the number of *A*-units which may be connected to a given sensory point.

(2) For simplicity of discussion, we will assume a single binary *R*-unit, to which every *A*-unit has an output connection. The connection of the *i*th *A*-unit to *R* has a weight associated with it, which is designated w_{ir} . This weight may be either positive or negative, according to the history of the unit, and is functionally equivalent to the "value," v_i , which has been assumed in previous analyses.^(1,2,3)

(3) Every *A*-unit receives n_x input connections from other *A*-units. Each of these connections may originate from any one of the other *A*-units, chosen at random. There are no constraints upon the number of cross-connections which may originate from any one *A*-unit, but the expected value of this number will also be n_x . A cross-connection which originates from unit a_i and terminates on

unit a_j has associated with it a weight, w_{ij} , which may be either positive or negative, depending upon the past history of the system. The weight, w_{ij} , is a measure of the signal transmitted from a_i to a_j when a_i is active.

(4) Every A -unit has a threshold, θ , which is assumed to be constant and equal for all A -units.

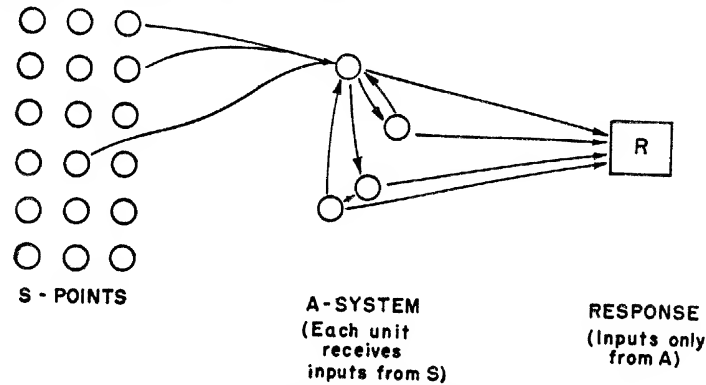


FIG. 4. Typical connections for a cross-coupled perceptron.

All signals transmitted in the perceptron will be measured in units which can be related to the threshold, θ . These units of signal strength, or intensity, will be called *standard threshold units*, or s.t.u.'s. An s.t.u. is of arbitrary magnitude relative to physical variables, but sets an internally consistent scale for the perceptron. The threshold, θ , is typically assigned a value of several s.t.u.'s, while sensory signals are either $+1$ or -1 s.t.u. The same unit will be employed as a measure of the weights w_{ij} and w_{ir} .

The rules for the dynamics of the cross-coupled perceptron are as follows:

(1) If the total input signal to an A -unit, which will be called α , is equal to or greater than the threshold, θ , the unit is active. The total input signal consists of the algebraic sum of positive and negative sensory signals, and the weights of any cross-connections from active A -units, as well as any feedback signal which a previously active A -unit may transmit to itself (see below). We will define the *activity coefficient* of unit a_i at time t by:

$$\alpha_i^*(t) = \begin{cases} 1 & \text{if } \alpha_i(t) \geq \theta \\ 0 & \text{if } \alpha_i(t) < \theta \end{cases}$$

(2) Each A -unit which is active at time t delivers a signal to all units to which it is connected; this signal arrives at the receiving units at time $t + 1$. The signal from a_i to a_j is proportional to w_{ij} .†

(3) Let U represent either an A -unit or an R -unit which is active at time t . Then the set of connections which terminate on this unit will have their weights permanently incremented or decremented according to the following rule: All connection weights, w_{iu} , which originate from a unit, a_i , which is also active at time t , gain an increment, Δw . At the same time, the weights of *all* connections to U lose a compensating quantity, equal to $n_x^* \Delta w / n_x$, where n_x^* = the number of connections which are "active" (i.e. which originate from active units) at time t . Thus, for an active connection, the net change is

$$\Delta w - n_x^* \Delta w / n_x$$

while, for an inactive connection, the net change is $-n_x^* \Delta w / n_x$. This rule (which makes this a "gamma system") guarantees that the sum of the input weights to any A -unit or R -unit will be zero at all times. Note that an R -unit is dynamically equivalent to an A -unit, and differs from an A -unit only in the topological fact that it has no direct input connections from any sensory points.

(4) An R -unit responds in the same manner as an A -unit with a zero threshold. It is active at time t if the total weight of its inputs at time t is greater than zero; otherwise it is inactive. (We can assume that if the net input weight is exactly zero, there is a 50:50 probability of the R -unit being active or inactive.)

These seven rules (three for the topological organization of the network and four for its dynamics) completely specify the kind of perceptron which we are interested in analyzing. We will designate this type of system a *cross-coupled gamma system*.

For this class of systems, we now assert the following five propositions, where N_s is the number of sensory points, and N_A , the number of A -units, is assumed to be large. The notation $R(S)$ will be used to denote the response evoked by stimulus S .

† An alternative rule to dynamic rule No. 2 is: "Each A -unit which is active at time t delivers an auto-feedback signal to itself at time $t + 1$." A model which employs this rule has been analyzed, and is found to yield essentially identical phenomena, but with poorer performance, than the system considered here. A possible justification for introducing such an auto-feedback signal in a biological model may be found in Burns' proposed mechanism for the repetitive firing of neurones, described in Ref. 11.

FIVE PROPOSITIONS FOR THE CROSS-COUPLED SYSTEM

$R[S_x] \neq R[S_y]$

S represents any random-pattern stimulus.

$\doteq$ means "identical with a probability greater than chance."

For finite N_s

1. If T is any 1:1 transformation, and the perceptron is exposed to the sequence $S_1, T(S_1), S_2, T(S_2), \dots, S_n, T(S_n)$, then

$$R[T^{-1}(S_x)] \doteq R[S_x] \text{ and } R[T^{-1}(S_y)] \doteq R[S_y].$$

2. For the same conditions, $R[T^k(S_x)] \doteq R[S_x]$ and $R[T^k(S_y)] \doteq R[S_y]$.

3. Let G be any 1:1 transformation group and let

$$\mathcal{S}_i \equiv \{T_{i_1}(S_i), T_{i_2}(S_i), \dots, T_{i_n}(S_i)\} \dots T_{i_1}, T_{i_2}, \dots, T_{i_n}, \text{ and } T_g \in G.$$

Then, given the sequence $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_m$,

$$R[T_g(S_x)] \doteq R[S_x] \text{ and } R[T_g(S_y)] \doteq R[S_y].$$

For $N_s = \infty$.

4. If S 's are random-pattern stimuli, all of the above biases vanish.
5. If S 's are such that expected illumination density in the neighborhood of an illuminated point $>$ expected illumination density over the retina, response biases in 1, 2, 3 should occur.

In the proof of these propositions, three functions, which will be called P_u , $P(k)$, and $P_x(k)$, are of fundamental importance. In what follows, we shall first carry out an analysis of these three functions, and then attempt to demonstrate that the above propositions necessarily follow.

The function P_u will be considered first. It is assumed that some stimulus, $S(t)$ occurs at time t , followed by $S(t+1)$ at time $t+1$. In the association system, $S(t)$ activates a set of A -units the measure of which has the expected value P_{a1} . $S(t+1)$ will activate a similar set of A -units, the measure of which is P_{a2} . The expected proportion of A -units responding to *both* stimuli is equal to the intersection of these sets, which has an expected value P_{a12} . These P_a functions (as functions of the threshold, θ) have been analyzed previously^(1,2) and always take the form of a monotonically decreasing function of θ .

P_u is then defined by the expression:

$$P_{u12} = P_{a1} - P_{a12} \quad (3)$$

i.e., P_u is the measure of the complement of the intersection ($P_{a_{12}}$) on the set responding to $S(t)$.

P_u will be equal to zero if the stimuli are identical, and will be greater than zero otherwise. If the intersection of the two stimuli (measured on the retina by the normalized "common" area, C_{12}) is equal to its expected value, $P_{a_{ij}} = P_{a_i}P_{a_j}$.

Since, in what follows, we will generally be concerned with random pairs of random-pattern stimuli (i.e. patterns of dots randomly scattered over the retina) the intersections will not vary greatly from EC_{ij} . Moreover, if all stimuli are of the same area (as is the case in a 1:1 transformation) this further simplifies to:

$$P_u = P_a - P_a^2 \quad (4)$$

Before proceeding to the remaining two functions, $P(k)$ and $P_s(k)$, we must define a concept of *similarity* for a pair of A -units, a_i and a_j . We will use the notation: $a_i \text{ sim } a_j (k, T)$ to mean that a_i has exactly k sensory origin points which are images under the transformation T of origins of a_j . For an origin point to qualify as an image under T , it is necessary not only that its location must be the T -transform of the location of some origin point of a_j , but that the connections in question should agree in sign as well; i.e. the origin point of an inhibitory connection can have an image only among the origin points of inhibitory connections of the other A -unit, and excitatory origins can have only excitatory images. If $a_i \text{ sim } a_j (k, T)$ we can also say that " a_i is similar to a_j by the criterion (k, T) ." Note that if $a_i \text{ sim } a_j (k, T)$, then $a_j \text{ sim } a_i (k, T^{-1})$, where T^{-1} is the inverse transformation.

We define $P(k)$ as the probability that two A -units have exactly k similar origins, under the transformation T . As long as T is a 1:1 transformation, $P(k)$ is independent of T , and is given by the expression:

$$P(k) = \text{Prob } \{a_i \text{ sim } a_j (k, T)\} = \sum_{k_e \text{ min}}^{k_e \text{ max}} P(k_e) \cdot P(k_i) \quad (5)$$

where $k_i = k - k_e$

$$P(k_e) = \binom{n_e}{k_e} P_e^{k_e} (1 - P_e)^{n_e - k_e}$$

$$P(k_i) = \binom{n_i}{k_i} P_i^{k_i} (1 - P_i)^{n_i - k_i}$$

$$k_e \min = \max(0, k - n_i)$$

$$k_e \max = \min(k, n_e)$$

$P(k_e)$ is the probability that exactly k_e out of the n_e excitatory connections to a_i are images of excitatory origins of a_j , and, similarly, $P(k_i)$ is the probability that k_i of the inhibitory origins of a_i are images of inhibitory origins of a_j . To calculate these

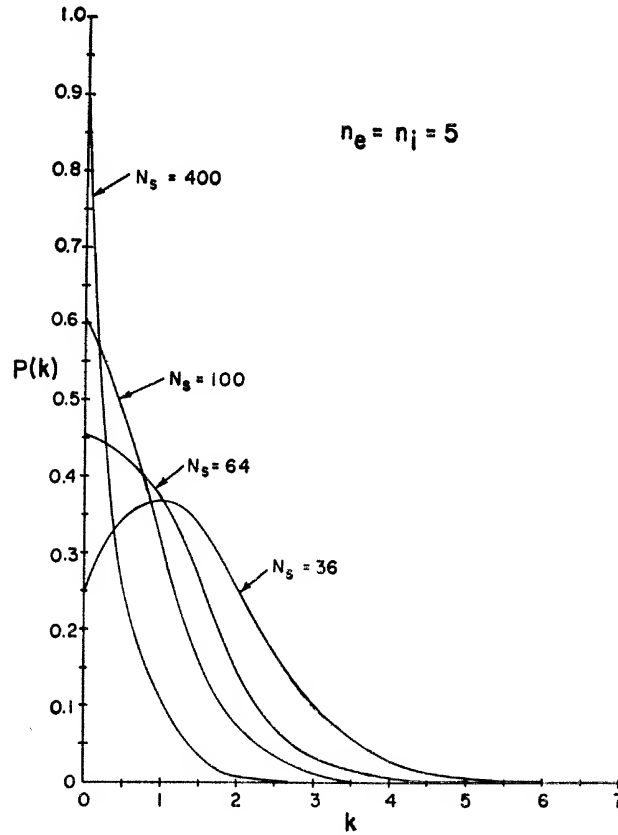


FIG. 5. $P(k)$ as function of N_e .

probabilities, we need the probabilities P_e and P_i , which are the probabilities that one of the excitatory or inhibitory origins, respectively, corresponds in location to an image point of a suitable

excitatory or inhibitory origin of a_j . These quantities are given by

$$P_e = 1 - (1 - 1/N_s)^{n_e}$$

and

$$P_i = 1 - (1 - 1/N_s)^{n_i}$$

where N_s is the number of sensory points in the retina. It follows from eq. (5) that $P(k > 0)$ goes to zero as N_s goes to infinity. A set of representative curves for $P(k)$ for a system with 5 inhibitory and 5 excitatory connections to each A -unit is shown in Fig. 5.

We now require the function $P_{*}(k)$. If we are given two units, a_i and a_j , which are similar by criterion (k, T) , then $P_{*}(k)$ is the probability that a_i responds to $T(S)$ given that a_j responds to S . If it is desired to indicate the particular units in question, and to emphasize the fact that this is a conditional probability, we may use the notation $P_{*i|j}(k)$, as distinct from $P_{*i,j}(k)$ which is used to denote the unconditional probability that a_i responds to $T(S)$ and a_j responds to (S) . Since the probability that a_j responds to S is P_{a_j} , we have

$$P_{*i|j}(k) = \frac{1}{P_{a_j}} \cdot P_{*i,j}(k) \quad (6)$$

We will therefore concentrate on finding the equation for $P_{*i,j}(k)$.

The set of active connections which carry impulses from the retina to a_i and a_j in the presence of stimuli $T(S)$ and S , respectively, can be divided into six independent subsets, containing $n_1, n_2, \dots, n_6$ elements, respectively, as follows:

- n_1 = number of a_i excitatory origins activated by $T(S)$ which are not images of a_j origins.
 - n_2 = number of a_i inhibitory origins activated by $T(S)$ which are not images of a_j origins.
 - n_3 = number of a_i excitatory origins activated by $T(S)$ which are images of a_j origins = number of a_j excitatory origins activated by S which are images of a_i origins.
 - n_4 = number of a_i inhibitory origins activated by $T(S)$ which are images of a_j origins = number of a_j inhibitory origins activated by S which are images of a_i origins.
 - n_5 = number of a_j excitatory origins activated by S which are not images of a_i origins.
 - n_6 = number of a_j inhibitory origins activated by S which are not images of a_i origins.
-

For a_i to be activated by $T(S)$ we require $n_1 - n_2 + n_3 - n_4 \geq \theta$. For a_j to be activated by S we require $n_3 - n_4 + n_5 - n_6 \geq \theta$. k_e, k_i, n_e, n_i and m are defined as before. The normalized area of the stimulus (i.e. the fraction of the retinal points which are illuminated) is equal to R , which, under the 1:1 rule is necessarily identical for S and $T(S)$. $k_i = k - k_e$. The conditional probability of k_e given k is

$$P(k_e | k) = P(k_e, k_i | k) = \frac{P(k_e) \cdot P(k_i)}{P(k)} \quad (7)$$

which can be calculated from eq. (5).

Going to the limit in N_s , we obtain:

$$\lim_{n_s \rightarrow \infty} P(k_e | k) = \frac{\binom{n_e}{k_e} \binom{n_i}{k_i} n_e^{k_e} n_i^{k_i}}{\sum_{k_e \min}^{k_e \max} \binom{n_e}{k_e} \binom{n_i}{k_i} n_e^{k_e} n_i^{k_i}}$$

Given k_e , we can write the conditional probabilities for our six n 's as follows:

$$\begin{aligned} P(n_1 | k_e) &= \binom{n_e - k_e}{n_1} R_1^{n_1} (1 - R_1)^{n_e - k_e - n_1} \\ P(n_2 | k_e) &= \binom{n_i - k_i}{n_2} R_2^{n_2} (1 - R_2)^{n_i - k_i - n_2} \\ P(n_3 | k_e) &= \binom{k_e}{n_3} R_3^{n_3} (1 - R_3)^{k_e - n_3} \\ P(n_4 | k_e) &= \binom{k_i}{n_4} R_4^{n_4} (1 - R_4)^{k_i - n_4} \\ P(n_5 | k_e) &= \binom{n_e - k_e}{n_5} R_5^{n_5} (1 - R_5)^{n_e - k_e - n_5} \\ P(n_6 | k_e) &= \binom{n_i - k_i}{n_6} R_6^{n_6} (1 - R_6)^{n_i - k_i - n_6} \end{aligned} \quad (8)$$

For an infinite retina, $R_1 = R_2 = \dots = R_6 = R$. The product of these six probabilities gives us the probability of a particular combination, $n_1, n_2, \dots, n_6$, given k and k_e . Summing these products

for all admissible combinations of $n_1 \dots n_6$ yields the probability that both A -units are activated by their appropriate stimuli, given k and k_e . To eliminate the conditional dependence of this result on k_e we multiply by $P(k_e | k)$, and sum the resulting (unconditional) probability over all admissible values of k_e , which gives us the value of $P_{x_{i,j}}(k)$. Performing all of these operations thus yields the following equation:

$$P_{x_{i,j}}(k) = \sum_{k_e \min}^{k_e \max} P(k_e | k) \cdot \sum_{\substack{n_1 - n_2 + n_3 - n_4 \geq \theta \\ n_3 - n_4 + n_5 - n_6 \geq \theta \\ n_1, n_5 \leq n_e - k_e \\ n_2, n_6 \leq n_i - k_i \\ n_3 \leq k_e, n_4 \leq k_i}} \prod_{v=1}^6 P(n_v | k_e) \quad (9)$$

A set of curves for the conditional form of $P_x(k)$, see eq. (6), is shown in Fig. 6, for the same values of n_e and n_i which were used in Fig. 5. Note that when $k = m$, the conditional probability, $P_{x_{i,j}}$, is equal to 1.

It should now be noted, however, that the above equations, and the curves shown in Fig. 6, are good only for the assumption of an infinite retina, in which case the only admissible value of k is 0. If the retina is finite, then the specification of one of the components, say n_1 , sets the constraint that the origin points from which this set of connections originates cannot be used again as members of a mutually exclusive set, such as n_3 . The mutually exclusive sets are: n_1, n_3 and n_5 ; n_2, n_4 and n_6 . Although we have placed no constraints upon the number of connections which may originate from the same origin point, the above combinations are mutually exclusive since if a point is employed in n_1 , this implies that it does *not* have an image among the a_j origins, while any point employed for n_3 does have such an image. n_3 and n_5 are similarly mutually exclusive. The relationship between the n_1 set and the n_5 set is less direct. If points are specified for n_1 these points are not themselves prohibited as origins for the n_5 set, but they imply the existence of an equal number of image points, which *are* excluded, for if any member of the n_5 set originated from one of these image points, it would have to be counted in n_3 instead of n_5 . The same observation

applies to the set of even-numbered n 's. For a very large, or infinite retina, the restriction of a small number of points by the above considerations will have no effect upon the probability that some other point is illuminated by the stimulus, and consequently $R_1, R_2, \dots, R_6$ are all equal to R in eq. (8). In a finite retina, however, the specifications of any of these components effectively reduces the residual set of illuminated points, and this causes a change in the value of R for the remaining components. Since origin points do not qualify as images, under our definition, unless the connections are of the same sign, the excitatory and inhibitory components do not interfere with one another, and remain independent. If we assume that none of the connections in any one of the first four components originate from the same point (which is a safe approximation in all but the smallest of retinas) we can estimate suitably corrected values for the R 's by assuming the following procedure:

- (1) Lay down k_e excitatory and k_i inhibitory origins, constituting the set of k imaged origins. Assuming no coincidences among the k_e excitatory origins, and no coincidences among the k_i inhibitory origins, this leaves $N_s - k_e$ sensory points as admissible origins for the n_1 component, and $N_s - k_i$ sensory points as admissible origins for the n_2 component. The probability that any one of the k imaged origins falls within the illuminated area of the stimulus is equal to $R = R_3 = R_4$. Having established these origin points gives us

$$R_1 = \frac{RN_s - n_3}{N_s - k_e}$$

$$R_2 = \frac{RN_s - n_4}{N_s - k_i}$$

- (2) Now assign the remaining origins of a_i , and count those points which fall within the admissible illuminated area (with probability R_1 or R_2) to obtain n_1 and n_2 . Since n_1, n_3 and n_5 are mutually exclusive, this leaves $RN_s - n_1 - n_3$ admissible sensory points as possible n_5 origins, and similarly we are left with $RN_s - n_2 - n_4$ admissible points for n_6 origins, while the total number of admissible points in the retina for excitatory origins has been reduced

to $N_s - n_e$, and for inhibitory origins has been reduced to $N_s - n_i$. This yields:

$$R_6 = \frac{RN_s - n_1 - n_3}{N_s - n_e}$$

$$R_6 = \frac{RN_s - n_2 - n_4}{N_s - N_i}$$

Substituting these values in eq. (8) yields a corrected expression for $P_x(k)$, for a finite retina.

On the scale used in Fig. 6, the difference between the finite and infinite cases would hardly be noticeable, provided the retina contained, say, 100 or more elements. Nonetheless, the difference is of considerable theoretical importance, as we shall soon see. It can be examined to better advantage in Fig. 7, which shows the difference, $D = P_x(k) - P_a$, for a number of finite retinal sizes, compared with an infinite retina. If the difference, D , for different values of k is weighted by $P(k)$ and summed, it should come out

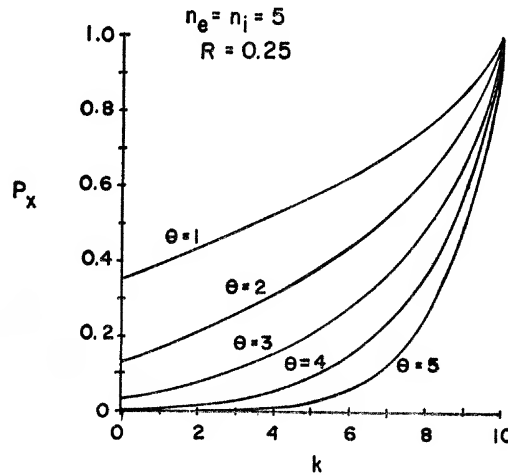


FIG. 6. $P_{x_i|j}$ as function of k for infinite retina.

to zero; i.e. the mean value of $P_x(k) = P_a$. As a consequence, we note that for a finite retina, D is actually negative for small values of k , although it converges to the same limit as the infinite case

$(1 - P_a)$ as k approaches m . For the infinite retina, D is equal to zero for $k = 0$, and greater than zero everywhere else—a fact which is consistent with the requirement that the sum of the weighted D 's must be zero if we recall that when $N_s = \infty$, $P(k) = 0$ for all values of k greater than zero.

We are now ready to consider the first of our propositions, namely:

PROPOSITION 1

If N_s is finite, and for two stimuli, S_x and S_y , we have $R(S_x) \neq R(S_y)$, and the perceptron is exposed to the sequence of random-pattern stimuli: $S_1, T(S_1), S_2, T(S_2), \dots, S_n, T(S_n)$, where T is any 1:1 transformation, it is predicted that upon presentation of $T^{-1}(S_x)$ and $T^{-1}(S_y)$ we will obtain the responses $R[T^{-1}(S_x)] \doteq R(S_x)$, and $R[T^{-1}(S_y)] \doteq R(S_y)$.

We recall that the sign $\doteq$ is interpreted to mean "identical with a probability greater than chance." Also, note that none of the stimuli $S_x, S_y, T(S_x)$ or $T(S_y)$ are assumed to occur in the course of the "pre-conditioning series" $S_1, T(S_1)$, etc. The significance of this proposition is that it asserts that (assuming proper parametric conditions) we can show our perceptron a series of random dot stimuli alternating with their transformations, $T(S)$, and expect to find a bias to generalize a response which is associated to any particular stimulus, S_x , so that it is now elicited with a greater than chance probability by $T^{-1}(S_x)$, while an opposite response which is associated to S_y generalizes similarly to $T^{-1}(S_y)$. It makes no difference whether the responses $R(S_x)$ and $R(S_y)$ have been learned before or after the preconditioning series. If this proposition is correct, it means, in effect, that we can expect *instantaneous generalization* from a stimulus to its transform, provided only that the perceptron has experienced a sufficient number of completely irrelevant stimuli to which the same transformation was applied.

Let us consider how this proposition arises from the original seven rules by which our system was defined. We will begin with an analysis of the expected bias which is introduced by showing the perceptron the simplest preconditioning sequence: the random-pattern stimulus S , followed by $T(S)$.

Let us first satisfy ourselves that the first stimulus, S , presented alone, introduces no bias of any consequence. The bias which we shall be looking for is a tendency for units which are similar by

criterion (k, T) to become more strongly coupled if k is large than if k is small. The measure of "coupling" between the A -units in question is the connection weight, w_{ij} . If we can show that the expected value of this weight is the same after the presentation of the random stimulus S as before, regardless of k , then it will become apparent later that the generalization bias predicted by the proposition cannot be obtained in this fashion.

Now, when the stimulus S is presented, a proportion P_a of the A -units is activated, which may include pairs of units having any degree of similarity, as measured by k . Since the points which comprise the random stimulus S bear no necessary relationship to the transformation T , however, there is no bias which favors one value of k more than another, among the various pairs of active A -units. Consequently, we expect pairs of A -units with $k = 0, k = 1, k = 2$, etc. to be represented among the active units in the same proportion as exists in the population of A -units at large, and any change in the connections of active units should be entirely independent of k . Moreover, the rules of the gamma system, defined previously, require that the net change in the weights of the cross-connections to any A -unit must be zero. Consequently, the expected value of the change (Δw_{ij}) in any cross-connection will likewise be zero, and (since the expected change is independent of k) the expected change in the weight of any pair of A -units which are similar by (k, T) is still zero, regardless of the value of k . This proves our assertion that the presentation of S alone introduces no coupling-bias which distinguishes between large and small values of k .

Now let us consider what happens when the second stimulus, $T(S)$, is presented immediately following the stimulus S . We have seen in Fig. 5 and the discussion of P_u , that the set of units activated by the second of two stimuli which follow in immediate succession includes a greater-than-normal complement of those units which responded to the first stimulus. In the present application, this means that $T(S)$ will activate an excess, measured by P_u , of units which normally respond to S . We might expect that this would introduce a bias in the population of active units favoring those pairs having a high degree of similarity under T ; i.e., we would expect to find an excess of active pairs, $a_j \text{ sim } a_i(k, T)$ where k is large, and a corresponding shortage of pairs whose similarity,

k , is small. If this is the case, we would expect the coupling of highly similar pairs to gain in value, while (through the gamma system compensation effect) connections of pairs which have a low value of k would be more likely to lose in value. That this is indeed the case we shall see through the following quantitative analysis.

Let S occur at time $t - 1$ and $T(S)$ at time t . (T is any 1:1 transformation). All A -unit connections, with their associated weights, w_{ij} , are divided into four classes based on the activity of the terminal units, a_i and a_j , as follows:

- Class 1: a_i and a_j both active at t
- Class 2: a_j active and a_i inactive at t
- Class 3: both members inactive at t
- Class 4: a_i active and a_j inactive at t

We will assume the convention that, in each case, the connection *originates* on the unit a_i and *terminates* on a_j . Then the effect of reinforcing the cross-connections at time t will be:

- Over all Class 1 connections, $E\Delta w_{ij} > 0$
- Over all Class 2 connections, $E\Delta w_{ij} < 0$
- Over all Class 3 connections, $E\Delta w_{ij} = 0$
- Over all Class 4 connections, $E\Delta w_{ij} = 0$

Note that these expected changes in the connection weights are independent of k . Let $E\Delta w_{ij}$ over the set of Class 1 connections $= \Delta_1$, let $E\Delta w_{ij}$ over the set of Class 2 connections $= \Delta_2$. We will now compute the expected gain in weight for a w_{ij} linkage between a pair of units of known similarity, k , as a result of exposure to our sequence $S, T(S)$. That is, we require $E\Delta w_{ij} | k$, where it is assumed that a reinforcement increases "active-active" connections by Δw , and at the same time reduces *all* connections terminating on an active unit by $-n_x^* \Delta w / n_x$, in accordance with our dynamic rule (3). But the expected value at time t of $n_x^* / n_x = P_a(t)$, so that the expected loss component will be $-P_a(t) \cdot \Delta w$. This means that

$$\begin{aligned}\Delta_1 &= \Delta w - P_a(t) \Delta w = [1 - P_a(t)] \Delta w \\ \Delta_2 &= -P_a(t) \Delta w\end{aligned}\tag{10}$$

From our discussion of P_u , it is clear that $P_a(t)$ will be equal to $P_{a_s} + P_u$, if P_{a_s} is the "normal" value of P_a in response to S , in the absence of feedback. Since these deltas are independent of k , we

can apply them to the set of connections for those pairs of A -units having any given value of k . Let

$N(k, T)$ = the number of connections between A -units which are similar by (k, T)

$N_1(k, T)$ = the number of connections between units which are similar by (k, T) and in Class 1

$N_2(k, T)$ = the number of connections between units which are similar by (k, T) and in Class 2.

Then we have:

$$E\Delta w_{ij} | k = \frac{\Delta_1 \cdot N_1(k, T) + \Delta_2 \cdot N_2(k, T)}{N(k, T)} \quad (11)$$

If we divide each term of the numerator and denominator by N_{ij} = the total number of a_i, a_j connections, the N 's will be converted into probabilities, as follows:

$$E\Delta w_{ij} | k = \frac{\Delta_1 \cdot P_1 + \Delta_2 \cdot P_2}{P(k)} \quad (12)$$

where $P(k) \equiv$ probability that $a_i \text{ sim } a_j(k, T)$

$P_1 \equiv$ probability that a connected (a_i, a_j) pair is similar by (k, T) and is in Class 1

$P_2 \equiv$ probability that a connected (a_i, a_j) pair is similar by (k, T) and is in Class 2.

Dividing the last expression through by $P(k)$ to obtain conditional probabilities, we obtain:

$$\begin{aligned} E\Delta w_{ij} | k &= \Delta_1(P_1 | k) + \Delta_2(P_2 | k) \\ &= \{ [1 - P_a(t)](P_1 | k) - P_a(t)(P_2 | k) \} \cdot \Delta w \end{aligned} \quad (13)$$

where $P_a(t) = P_{a_s} + P_{a_u}$, as before.

Let us now try to compute these conditional probabilities, $P_1 | k$ and $P_2 | k$. We recall that for Class 1 connections, a_i and a_j must both be active at time t , while for Class 2, a_j must be active, and a_i inactive at time t . For the conditional probabilities, we further require that $a_i \text{ sim } a_j(k, T)$. If this condition is fulfilled, then the first probability, $P_1 | k$, will be the sum of the following four components. To simplify notation, we will let $P_{a_1} = P_{a_s}$, $P_{a_2} = P_{a_{T(S)}}$, and $P_{a_{12}} = P_{a_{sT(S)}}$.

- (1) Probability that a_j is in the set "normally" responding to $T(S)$ and a_i is in the set normally responding to $T(S) = P_{a_2}^2$.
- (2) Probability that a_j is in the set normally responding to S but not to $T(S)$, and a_i is in the set normally responding to S but not to $T(S) = P_u^2$.
- (3) Probability that a_j is in the set normally responding to S but not to $T(S)$, and a_i is in the set normally responding to $T(S) = P_u \cdot P_x$.
- (4) Probability that a_j is in the set normally responding to $T(S)$, and a_i is in the set normally responding to S but not to $T(S) = P_{a_2} \cdot P_u$.

Summing these four components yields:

$$P_1 | k = P_{a_2}^2 + P_u^2 + P_u P_x(k) + P_{a_2} P_u \quad (14)$$

A similar analysis can be performed for $P_2 | k$, which is the probability of having a_j active and a_i inactive, given k . Now a_j , the active member of the pair, must either be in the subset of units normally responding to $T(S)$ or else in the incremental " P_u set," normally responding to S but not to $T(S)$. Let a_j be in the " P_u set." Then the probability that a_j is in the P_u set and a_i is active is the sum of components (2) and (3) above $= P_u^2 + P_u P_x(k) = P_u[P_u + P_x(k)]$. The conditional probability of a_i active given a_j active is then $P_u + P_x(k)$. Therefore the probability that a_j is in the P_u set and a_i is inactive $= P_u(1 - P_u - P_x(k))$. Similarly, we obtain the probability that a_j is in the " P_{a_2} set" and a_i is inactive, which turns out to be $P_{a_2}(1 - P_u - P_{a_2})$. Summing these two components yields:

$$\begin{aligned} P_2 | k &= P_u[1 - P_u - P_x(k)] + P_{a_2}(1 - P_u - P_{a_2}) \\ &= P_u + P_{a_2} - P_u^2 - P_{a_2}^2 - P_u P_x(k) - P_{a_2} P_u \end{aligned} \quad (15)$$

Substituting for $P_1 | k$ and $P_2 | k$ in eq. (13), setting $P_{a_2} = P_{a_s}$, and dividing through by Δw gives us:

$$\begin{aligned} \frac{1}{\Delta w} \cdot E\Delta w_{ij} | k &= [1 - P_a(t)](P_{a_s}^2 + P_u^2 + P_u P_x(k) + P_{a_s} P_u) - \\ &\quad - P_a(t) \cdot (P_u + P_{a_s} - P_u^2 - P_{a_s}^2 - P_u P_x(k) - P_{a_s} P_u) \\ &= (P_{a_s}^2 + P_u^2 + P_u P_x(k) + P_{a_s} P_u) + (P_{a_s} + P_u)^2 \end{aligned} \quad (16)$$

Now let us substitute $P_x(k) = P_{a_s} + D(k)$, where D is defined as $P_x(k) - P_{a_s}$, as in Fig. 7. Thus, we finally obtain:

$$\begin{aligned} \frac{1}{\Delta w} \cdot E\Delta w_{ij} | k &= [P_{a_s}^2 + P_u^2 + 2P_u P_{a_s} + P_u D(k)] - (P_{a_s} + P_u)^2 \\ &= P_u D(k) \end{aligned} \quad (17)$$

In other words, we have proven that $E\Delta w_{ij} | k$ is proportional to the difference, $D(k) = P_x(k) - P_{a_s}$, and to the quantity P_u , which was previously defined. Now, still limiting ourselves to the

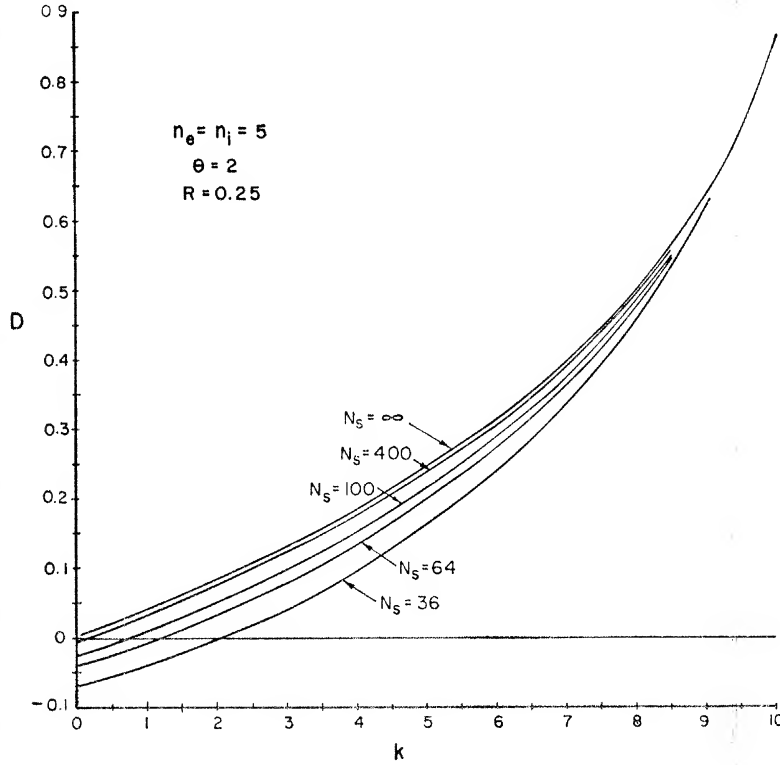


FIG. 7. $D = P_x - P_{a_s}$ for finite retina.

set of connections connecting units of similarity k , let us examine the bias which is introduced upon presentation of the stimuli $T^{-1}(S_x)$ and $T^{-1}(S_y)$, as specified in Proposition 1.

Either of these stimuli will activate a set of A -units which (as was seen in the discussion of the zero bias resulting from the presentation of a stimulus S alone) are not biased towards any particular values of k , either in their connections with other active units or in their connections with inactive units. Suppose $T^{-1}(S_x)$ is the stimulus presented. Then this will activate a set of n_x cross-connections, some of which terminate on units normally responding to S_x , while others terminate on units normally responding to S_y . The probability that one of these connections terminates on an A -unit which is similar by (k, T) to the first unit is equal to $P(k)$. If it is similar by (k, T) , then the probability that it is a member of the set responding to S_x is $P_x(k)$, while the probability that it is a member of the set responding to some other "random" stimulus, such as $T^{-1}(S_y)$ or S_y , is equal to P_a (being independent of k) in the case of any arbitrary set of A -units other than the set specifically responding to S_x . This means that if $T^{-1}(S_x)$ occurs, the probability that an output signal on a cross-connection between a pair of units which are similar by (k, T) terminates upon an A -unit in the set responding to S_x is equal to $P_x(k)$, while the probability that the output signal is transmitted to a unit in the set responding to S_y is equal to P_a . Now, we know that the weight of a cross-connection between two k -similar units has received an expected increment equal to $D(k)$. Therefore, the expected value of an input to a unit in the set normally responding to S_x through one of its cross-connections, when $T(S_x)$ is shown, will be $P_x(k) \cdot D(k)$. On the other hand, by similar reasoning, the expected value of the input received on a cross-connection by a unit in the set normally responding to S_y when $T(S_x)$ is shown, will be $P_{ax} \cdot D(k)$. The difference between the expected signal received by a unit in the S_x set and the expected signal received by a unit in the S_y set when $T(S_x)$ is shown, following the preconditioning series $S, T(S)$, will be:

$$\begin{aligned} E\Delta B | k &= (P_x(k) - P_{ax}) \cdot D(k) \Delta w P_u \\ &= D^2(k) \Delta w P_u \end{aligned} \quad (18)$$

The use of D^2 in the above equation assumes that all stimuli are of the same area, so that $P_{ax} = P_{ay}$. This is not a necessary condition (as can be seen by carrying along the two different values of P_a in the following analysis) but the equations are simplified by having a single value of P_a .

The above bias is the increment for each connection between a k -similar pair, favoring the excitation of an A -unit in the S_x set over a unit in the S_y set, when $T(S_x)$ is presented. The same argument yields an opposite bias when $T(S_y)$ is presented. We can now remove the dependence on k by averaging over all possible magnitudes of k as follows:

$$E\Delta B = P_u \Delta w \cdot \sum_k D^2(k) P(k) \quad (19)$$

This quantity is the *expected bias increment per connection per preconditioning stimulus pair*. Each combination, $[S, T(S)]$, in the preconditioning sequence will add this expected value to the bias for every connection, while the transitions $(T(S_1), S_2)$ which involve stimuli of different varieties, will contribute a zero increment, since all of the effects in the preceding analysis will be independent of k , and it can readily be shown algebraically that the expected bias increment is zero. Thus, if the preconditioning sequence consists of n_s random-pattern stimuli, alternating with their T -transforms, as required in Proposition 1, we get a total bias per A -unit of

$$\begin{aligned} \Delta B &= P_u n_s n_x E\Delta B \\ &= n_s n_x \Delta w P_u \cdot \sum_k D^2(k) P(k) \end{aligned} \quad (20)$$

This bias is actually equal to the expected total signal received from cross-connections by an A -unit in the S_x set as a result of presenting $T(S_x)$. An A -unit in the S_y set, under the same conditions, receives an expected total cross-connection signal of zero, as can readily be shown from the above equations.

Now, this means that if the expected bias, ΔB , becomes great enough (as a result of having seen many preconditioning stimuli) a unit in the S_x set will tend to be fired by $T(S_x)$, even though it is not normally in the intersection of the sets responding to S_x and $T(S_x)$. A unit in the S_y set, on the other hand, will not be so affected by a presentation of $T(S_x)$, but *will* be biased to respond to a presentation of $T(S_y)$. Consequently, a presentation of $T(S_x)$, in a large perceptron, will tend to activate more A -units normally responding to S_x than units normally responding to S_y , and, since it is these units which determine the response, $R(S_x)$ is more likely to occur than $R(S_y)$. Conversely, if $T(S_y)$ is presented, we expect

to activate more units in the set normally responding to S_y than in the set normally responding to S_x , so that $R(S_y)$ is more likely to occur than $R(S_x)$. This proves our Proposition 1.

In the above analysis, it has been assumed that the weights which gradually accumulate during the preconditioning cycle will not drastically alter the situation from the conditions prevailing for the first pair, $[S_1, T(S_1)]$. This will actually be so only if the increment per stimulus cycle, Δw , is small enough to prevent the accumulated weights from swamping the normal sensory input signals, and to keep the standard deviation $\sigma(w_{ij})$ of the distribution of connection weights small relative to the input signals. In other words, we would like to maintain $\sigma(w_{ij}) < 1$ s.t.u., while making the expected bias as large as possible by the end of the preconditioning series. If we set the conditions that

$$\begin{cases} \sigma(w_{ij}) = 1 \text{ s.t.u.} \\ EB = 1 \text{ s.t.u.} \end{cases}$$

at the end of the preconditioning series, then it is possible to solve these equations simultaneously for the required values of Δw and of n , the number of preconditioning stimuli. Making a rough approximation for the variance, this calculation was performed, and it was found that for a perceptron with 5 excitatory and 5 inhibitory connections per A -unit, and with 100 A -units each of which is coupled to the other 99, and using stimuli covering 25% of the retinal area in a 400-point retina, it would be necessary to use a Δw of 0.0001 s.t.u. and to show the perceptron on the order of 10^7 preconditioning stimuli. In other words, this is a very weak statistical effect compared to those which we have hitherto employed in the perceptron.

A better idea of the magnitudes of the expected bias under these conditions can be obtained from Fig. 8, which shows the expected bias increment per connection per preconditioning pair measured in units of Δw . Note that the bias has been multiplied by 10^4 for convenience of scale representation. The effect is strongly dependent upon the threshold, as well as upon the number of retinal points, N , against which it is plotted.

Attempts to simulate this effect upon the 704 computer are being made at the present time. The large number of cycles and the small magnitude of Δw required for a well behaved system, however,

make the problem exceptionally difficult, and no success has been achieved to date.*

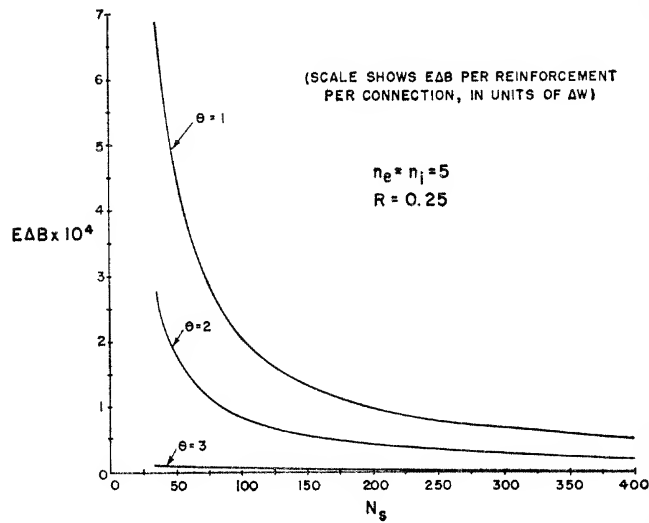


FIG. 8. $E\Delta B$ as function of N_s .

PROPOSITION 2

This states that after having seen the same preconditioning series discussed above, the perceptron will tend to generalize the response $R(S_x)$ not only to $T(S_x)$, but also to $T^2(S_x)$, $T^3(S_x)$, etc. That this should occur can be seen from the fact that the effect previously considered is based upon the activation of an additional set of A -units normally responding to $T^{-1}[T(S_x)]$. But, by extension of the same effect, this new set of units will tend to activate some set of A -units responding to $T^{-2}[T(S_x)]$, etc. This effect will admittedly attenuate rapidly as the power of T increases, but nonetheless it is theoretically present to some degree.

* Since this paper was written, the cross-coupled system has been successfully simulated, and the predicted effects observed. In order to keep the weights within acceptable bounds, an upper limit was introduced for w_{ij} , permitting greater values for Δw and more rapid growth of the bias than is indicated in the text.

PROPOSITION 3

This rests upon the fact that pairs of stimuli which are not related by a particular transformation, T , in general tend to introduce a zero bias, for any coupling in which T is involved. Consequently, the generalized sequences of transformations will not interfere with the development of biases favoring each of the particular transformations of the group, G , which is in question. Moreover, through the power-effect (Proposition 2) there should actually be some degree of mutual support between those transformations of G which are generated by the same power series.

PROPOSITION 4

This follows directly from the fact that $P(k) = 0$ everywhere except for $k = 0$, while $D = 0$ for the infinite retina when $k = 0$. Equation (20) yields a zero bias under these conditions.

PROPOSITION 5

Proposition 5 indicates the corrective for this condition, which would otherwise make this entire effect too small to be of practical use in any but the smallest of sensory fields. In dealing with a very large retina, or a retina of infinitesimal elements, it is necessary to take account of the *neighborhood* of an illuminated point and its transform in defining "similarity" for a pair of A -units. If we persist in using random-pattern stimuli (i.e. stimuli in which any retinal point is equally likely to be illuminated) there is indeed nothing that can be done about Proposition 4. Real-world stimuli, however, are not random assemblages of infinitesimal points; they tend to have some degree of *coherence*, and an origin point of a_j which does not actually originate from the ideal image-point of an origin of a_i may still be counted as "similar" if it is sufficiently *close* to the ideal image-point in question. Thus, in dealing with a universe of coherent stimuli, we define $a_i \text{ sim } a_j(k, d, T)$ to mean that a_i has exactly k origins which are within a distance d of image-points under T of origins of a_j . If we take the derivative of our various functions with respect to d , and ultimately integrate the derivative of the bias over d , we will find that the predicted effects are restored, since the integral of $P(k)$ over d is no longer equal to zero for $k > 0$. This analysis has not actually been carried out

quantitatively. Proposition 5 is based on the qualitative considerations above, but it is expected that the introduction of coherent stimuli in place of random point stimuli will markedly increase the magnitude of the bias effects, even for a relatively small retina.

V. CONCLUSIONS

The five propositions which we have just considered describe a number of hitherto unobserved phenomena in a fundamentally simple perceptron. This system is capable of "abstracting" those transformations which are most common in a particular environment, and applying them to new stimuli, which may be quite different in form from any which it has seen. It seems to accomplish all of the results of more rigidly designed systems, but arrives at its organization spontaneously, rather than having it built into the system. It is actually a system which "learns to learn," in the sense that prior to the preconditioning experience it would be able to generalize from a given stimulus to its transform only by the slow and laborious method of contiguity generalization, while after having seen a suitable preconditioning sequence (not including the stimuli to be used for test purposes), it performs the same task directly and without the requirement of any appreciable learning period. The concave curves which are characteristic of human learning in problems such as our horizontal-vertical bar experiment (Fig. 3) should now begin to appear. In fact, these curves should change progressively in the direction of a concave "insight learning" curves, as preconditioning experience accumulates. It might be interesting to check for a similar change in the character of perceptual generalization curves in biological experiments on young animals. It can also be shown that the problem of being able either to recognize the similarity or the dissimilarity of transformed stimuli, such as the dilemma of distinguishing the N from the Z in a system which generalizes over a group of rotations, has a satisfactory solution in this system.

It is hoped that future work will reveal the potentialities of these effects more clearly. Meanwhile, it seems most important to investigate possible methods of optimizing the bias effects, in order to make use of the weak statistical couplings upon which they are based.

REFERENCES

1. F. ROSENBLATT, *The Perceptron: A Theory of Statistical Separability in Cognitive Systems*, Cornell Aeronautical Laboratory Report No. VG-1196-G-1, Buffalo, January 1958.
2. F. ROSENBLATT, The perceptron, A probabilistic model for information storage and organization in the brain, *Psychol. Rev.* **65**, 386-408 (1958).
3. F. ROSENBLATT, Two theorems of statistical separability in the perceptron, *Proceedings of Symposium on the Mechanization of Thought Processes*, 1958. H.M. Stationery Office, London. Also printed without appendix as Cornell Aeronautical Laboratory Report No. VG-1196-G-2, Buffalo, September 1958.
4. R. L. GRIMSDALE, F. H. SUMNER, C. J. TUNIS and T. KILBURN, A system for the automatic recognition of patterns, *Proc. Inst. Elect. Engrs.* **106**, Part B (1959).
5. K. STEINBUCH, Automatische Zeichenerkennung, *Nachrichtentechnische Zeits.* **11**, 210-219 and 237-244 (1958).
6. O. SELFRIDGE, Pandemonium: A paradigm for learning, *Proceedings of Symposium on the Mechanization of Thought Processes*, 1958. H.M. Stationery Office, London.
7. D. O. HEBB, *The Organization of Behavior*, John Wiley, New York (1959).
8. J. T. CULBERTSON, *Consciousness and Behavior*, Wm. C. Brown, Dubuque, Iowa (1950).
9. W. PITTS and W. S. MCCULLOCH, How we know universals; the perception of auditory and visual forms, *Bull. Math. Biophysics* **9**, 127-147 (1947).
10. W. A. CLARK and B. G. FARLEY, Generalization of pattern recognition in a self-organizing system, *Proc. Western Joint Computer Conf.* 86-91 (1955).
11. B. DELISLE BURNS, The mechanism of after-bursts in cerebral cortex, *J. Physiology* **127**, 168-188 (1955).

DISCUSSION

KIRSCH (*Bureau of Standards*): In your description of what you call a group of one-one transformations over your retinal field, do I understand correctly that this set of transformations (which seem to do some radically disturbing things to this original letter N) all have the property that they preserve intact the area of the original figure—

ROSENBLATT: That is correct.

KIRSCH: In which case recognition over these transformation groups consists of nothing more than identification of a constant area for a figure?

ROSENBLATT: No. We are trying to apply a particular constrained group of possible transformations. We are not allowing all possible one-one transformations to appear in a given universe. For example, we are interested in a universe in which possibly only rotations occur, or possibly one particular movement of the stimulus. We then predict that this particular transformation will be learned and abstracted by the system and applied to new stimuli as they come in.

The fact that the area is constant, as a matter of fact, has nothing to do with it. Our first proposition asserts that opposite responses can still be elicited by different stimuli, even though they have the same area, provided the stimuli are not transformations of one another under the group in question.

QUESTION: You mentioned that the biological models of learning did not match the curve you had of behavior. I don't believe you were looking at the

right model of learning. When we do things where we don't know what we are doing, like in studies of taxonomy and in trying to catalog Chinese characters, we tend to start off lickety-split, we know just about what we are doing, and then pretty soon the fringe cases come about and we don't know quite where they fit in and it takes a long time to settle down. I think that is more appropriate than the pair-associated model.

ROSENBLATT: That is certainly true in certain kinds of learning problems. I don't think, however, it is fair to resort to this type of learning material or this type of experimental situation in drawing comparisons with something like separating horizontal bars from vertical bars. In this particular experiment, these curves seem to be wrong for an adult human observer. This is the point I was trying to make.

QUESTION: I would like to know what experimental evidence there is for the statement you made that different organisms have much different connections between say the sensory cells and the central nervous cells?

ROSENBLATT: The evidence I think begins as soon as you begin trying to identify areas in the brain of two members of the same species. It is hard enough even to be sure you are talking about exactly the same functional area of the brain, let alone being able to match connections on a one to one basis from one member of the species to another.

If you do pick an area to be identical, I challenge you to find two members of the species with all cells in this area connected in the same topological fashion. To a certain degree one can trace connections. One can certainly trace them to the degree where it becomes quite clear there is a great deal of variation of an apparently random sort between individuals.

MARTHA EVANS (*Los Alamos Scientific Laboratory*): You mentioned that the learning curves achieved in the horizontal-vertical discrimination were different from those of adult humans in similar learning situations. Are you attempting to duplicate human learning? Have you considered "better" modes of learning or would you?

ROSENBLATT: Well, first of all let me say that we are interested in duplicating human learning, if it is possible to do so. We are interested in determining the extent to which it is feasible to consider such a thing as duplicating human learning, or at least understanding how human learning operates. Whether or not there exists a better mode of learning is in a sense an empirical question to which I don't feel we can supply an answer at this point.

We are interested, however, not only in studying human learning, but in studying the behavior of networks which include biological nervous systems as a subclass. That is to say, we are interested in the study of signal transmission networks which involve connected nodes or cell points which have functional characteristics similar to those of biological neurons, but not necessarily identical.

If it emerges from the study of such systems that some of these behave better than others or some of them do in fact behave better than the human nervous system, this would be a very interesting finding indeed. But it would emerge from the study of this general class of systems and is not something I feel we can specifically aim for at this point.

KALIN (*Air Force Cambridge Research Center, Bedford, Massachusetts*): I would like to make clear, ladies and gentlemen, that I address this comment to Dr. Rosenblatt because his proposed machine is one of the better known examples of automaton type devices in which the connections over a large part of the system are essentially at random. I just want to raise a point which may have been overlooked concerning some of the implications of random connections

as sensory elements to nodes within the system. This has to do with arbitrary rearrangements of the retinal points on some two-dimensional retinal array on which we impose a pattern.

It does seem to me any arbitrary transposition of such retinal points leaves a machine the same as it was before. In other words, if we impose an arbitrary order on the outside of something randomly organized on the inside, we still have a randomly organized device on the inside and since any permutation is a product of transpositions—it seems we could arbitrarily rearrange these retinal elements in the same way and still have the same type of a machine.

From this point it seems to me that it necessarily follows that we can't define an adequate distance measuring function over the space of n points defined over this retinal array. In other words, the only metric we can define, satisfying the usual properties such that it is a positive number that is zero if the two points that are supposed to measure the distances between are identical and in other cases it satisfies the triangle inequality and things of this sort—that the only metric we can define which measures the distance between two points arbitrarily fixed at A and B, is such that it is a constant if the points are not identical.

Now if that is the case, I think we have trouble in defining machine characteristics that can be interpreted as meaning that the various points are adjacent to each other or that a certain set of points are equidistant from another point. In other words, a comparative metric problem. And this bothers me because it seems from this it would follow that we have to admit into consideration not just the class of patterns that the experimenter is interested in, systems which involve continuity, contiguity, and so forth, but all possible abstract classes of the bits that make up this retinal field, and there is an awfully large number of these.

For an n bit field, suppose we have a square of $\sqrt{n}$ on the side; we can make up 2^n arbitrary patterns and we are interested in classes of these patterns. There is an even larger number of those classes. There is 2^{2^n} . Now if we show a certain limited number of patterns, say P , it still remains, that there are $2^{2^n - P}$ —different classes in the abstract, all of which contain the initial frequency of T patterns; and if our automaton is randomly organized there is no reason why it should pick one of the members of those classes that make sense to us as opposed to any other.

I simply throw this question out. I think we have to get in there and monitor the thing somehow. We have got to tell it when it is doing wrong and encourage it when it is doing right as this learning trial proceeds. That is the gist of my comment.

ROSENBLATT: I think there are one or two mistaken assumptions which are fundamental to this argument. One is the idea that—and I think really I should refer back to Professor von Foerster's discussion this morning in this connection—this is the idea that if we have a random system in an organized environment, the randomness of the system will necessarily make all alternative organizations, after a period of experience in this environment, emerge with equal probability.

This is in general not the case. If we place a randomly organized system, particularly one which has a few constraints in it of the sort that we use in the perceptron organization (which has such constraints as the direction of the connections to each cell, and so forth) if we place a random network of this sort in an organized environment, then not all of the 2^{2^n} possible classes will emerge with equal probability. These are not equiprobable terminal states.

Now the question, and it is a very valid one, is whether the particular states in which the system is most likely to terminate are interesting ones. In some

cases they are not. And this is very much a function of the particular organized environment in question and the particular kind of system we are dealing with.

However, in a large number of cases we can guarantee that the classes are "interesting." For example, in the experiment which I illustrated this morning, in which a randomly organized network is placed in an environment of horizontal and vertical bars, we can guarantee that in every case the system, if it is organized in this particular fashion, will eventually form two classes, specifically the class of horizontal bars and the class of vertical bars. Regardless of how we were to extend these sequences into the future it would continue to respond consistently by placing all horizontal bars in one class and all vertical bars in another class.

Now these remarks and the question, I think, apply particularly to some of our former models in which we considered this question of spontaneous class formation in systems exposed to an environment of some specified type. This does not apply to the model we considered this morning, in which we assumed that there exists some association which might have been forced by the experimenter or might even have been built into the system. We then asked the question, given two responses which have already been assigned to two particular stimuli in an opposite sense, so that the stimulus S_x elicits one response and the stimulus S_y elicits another one—we now ask how these particular responses generalize over the various possible transformations of S_x and S_y . The only thing which is spontaneously organized here involves not so much the particular response, or the classes which are formed, but rather the ability to select out those transformations which have been applied, in the course of the systems' experience, to other stimuli, and then apply these same transformations to the particular stimuli in question, S_x and S_y , which have not formed part of the preconditioning history of this system.

Perhaps you would care to amplify your comments if you have anything further to say on this and maybe I am missing one of your points, but it seems to me that the essential thing on which your hypothesis rests, is the assumption of equiprobable terminal states of the system. If we show that we do not have equiprobable terminal states, then this argument is not really a particularly meaningful one, because out of the 2^n logical alternatives it is very possible that only one or two terminal states have a nonzero probability. Indeed this is the case in a number of systems which are nonetheless randomly organized.

KALIN: I would not like my comment to rest on the necessity for equiprobableness at this stage, but merely to point out that the probability of unwanted states may not be sufficiently close to zero to be convenient and I can't back this up by numbers at the moment. Really the gist of my comment, I think, is to ask us to keep in mind that a given set of sensory data can be organized in a very large number of ways and if the various ways in which it can be organized are not equiprobable, the particular classes in which we are interested might by some suitable measure be a little too small for convenience. I expect this effect to make itself apparent in the future as we develop models that have many alternative courses of action as opposed to just two alternative courses of action as we discussed here to-day.

ROSENBLATT: It is certainly possible that in dealing with systems with larger numbers of possible responses or categories we will start running into problems of this sort, and indeed I would expect that we would.

At that point we are going to get into realms in which the human subject would find it very difficult in many cases to decide which are the relevant classes to select, and admittedly a really equiprobable assumption is perhaps not essential to your argument; but the important thing is that the measure of the

probability of the particular few classes which we are interested in appears to be overwhelmingly strong in comparison to the measure of all of the other classes collectively.

TAUB (*Documentation Inc.*): I have a question that follows the question that Russell Kirsch asked earlier about patterns and size. We have heard a great deal and we will hear, I suppose, a great deal more about pattern recognition. I find the notion of pattern difficult to understand because they are all drawn in two dimensions on slides and I don't think those are real patterns in the real world.

I wanted to know, is a pattern made up of points? Can you take a line and break it down? Is each point a pattern or are two points patterns or three points patterns? Again, what effect is there on the notion of pattern of the fact that any area can be integrally divided? Does every pattern area contain an infinite number of patterns?

In other words, I would like to hear more discussion of the kinds of patterns that machines are supposed to recognize.

ROSENBLATT: In general, since this is apparently based purely on the visual model, let me again constrain my remarks to that. If we have a normal biological retina, this retina is a finite mosaic. It has a finite number of points and consequently the only possibility of conceiving of an infinite number of stimuli in terms of signals which could occur at the retina would be in terms of a continuous gradient of illumination at each one of these points. At this point I think we can readily resort to the information theory dodge of simply quantizing these continuous variables and recognizing that we can still represent everything impinging on the retina by a finite number of patterns.

THE ORGANIZATION AND REORGANIZATION OF EMBRYONIC CELLS*

ROBERT AUERBACH†

*Department of Zoology, University of Wisconsin,
Madison 6, Wisconsin*

LAST January I received from our Conference Chairman a list of titles of all the speakers for this meeting. It was indeed a frightening experience seeing things like "networks," "growing automata," and "strategies"; but being basically optimistic I reasoned that perhaps my interdisciplinary colleagues might feel equally indisposed upon seeing the topic "Growth and Differentiation of Isolated Embryonic Tissues in Response to Specific Inductive Stimuli." This, indeed, was the title I had submitted for my talk to-day. If "corrigible strategy of thinking" is what I think it is, I used this, called the Conference Chairman, and changed my title to be less specific and more ambiguous.

Anyway, I am an embryologist, so that no matter what the title is, I will be talking about embryos and about questions relating to developmental processes. Specifically, the investigations to be discussed concern the question of cellular organization and reorganization during embryonic development.

For an embryologist, basically interested in the processes by which the fertilized egg cell develops into the adult organism, there is no problem in finding organization to talk about; it is, rather, trying to limit the topic that is difficult. Organization is the crux of embryological investigation: The egg itself shows complex organization

* Previously unpublished work reported in this paper was supported by a research grant C-3985 from the National Cancer Institute, National Institutes of Health.

† With the technical assistance of William D. Ball.

at every level of analysis—in the distribution of particulate components, in the gradients of biochemical entities, in the structural configurations within each sub-unit, and so on. At fertilization, sperm entry leads to a complex pattern of reorganization, again at every level of analysis; and indeed each succeeding step is characterized by reorganization until at the other end of the embryological time scale the development of organs, organ systems and organisms literally shout organization in their very names. We will not consider either end of the scale, but will limit ourselves to discussing a few isolated differentiative events. Similarly we will consider only a single level of organization, that of tissue organization, in which the basic unit is the cell. Within these limited boundaries we will ask the questions: What are the inherent organizing properties? By what mechanisms do these cells establish and maintain organization? What is the effect of differentiative stimuli on these properties?

Two phenomena central to embryological thinking are regulation and induction—the first represents intrinsic organizing properties, the second extrinsic ones. We can define these two phenomena operationally as follows:

- (1) A limb bud is halved, or two limb buds are fused; a single, harmonious limb is formed. Regulation has occurred.
- (2) Gastrular ectoderm, isolated in tissue culture, will develop into epidermis. If such tissue is exposed to chorda-mesoderm tissue it does not become epidermis, but becomes nervous tissue instead. Induction has occurred.

Both these processes represent embryological communication. In the case of regulation, the cells of the regulating mass must be able to communicate with each other, to let each other know, as it were, where they are in relation to each other; for in some way these cells form a single, harmonious, complete structure.^(1,2) In the second case the communication is overtly simpler, since one group of cells “tells” another group to change; but again the change is to a new harmonious system. And, indeed, the communication during induction is usually a two-directional one, which is why in much of the ensuing discussion we will talk about inductive interaction. Bonner has recently used the term “chemical

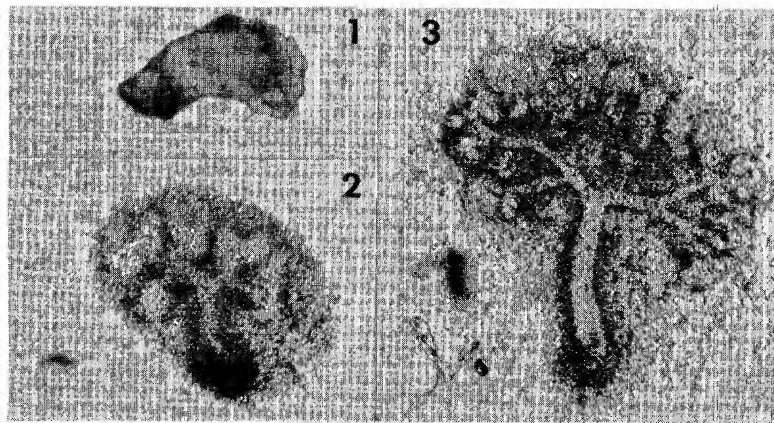


FIG. 1. Eleven-day mouse embryo metanephric kidney at time of isolation in tissue culture.

FIG. 2. Same kidney, after 2 days in culture.

FIG. 3. Same kidney, after 6 days in culture.

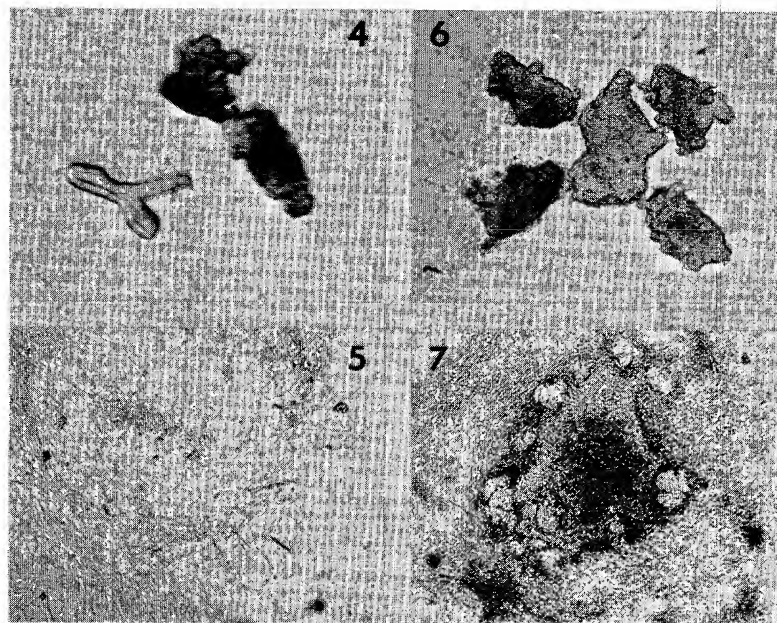


FIG. 4. Eleven-day kidney, separated into ureteric bud and mesenchyme.

FIG. 5. Absence of tubule formation by kidney mesenchyme after 3 days in culture in the absence of inductively active tissue.

FIG. 6. Typical combination of kidney mesenchyme with dorsal spinal cord.

FIG. 7. Tubule formation by kidney mesenchyme after 3 days in culture with dorsal spinal cord.

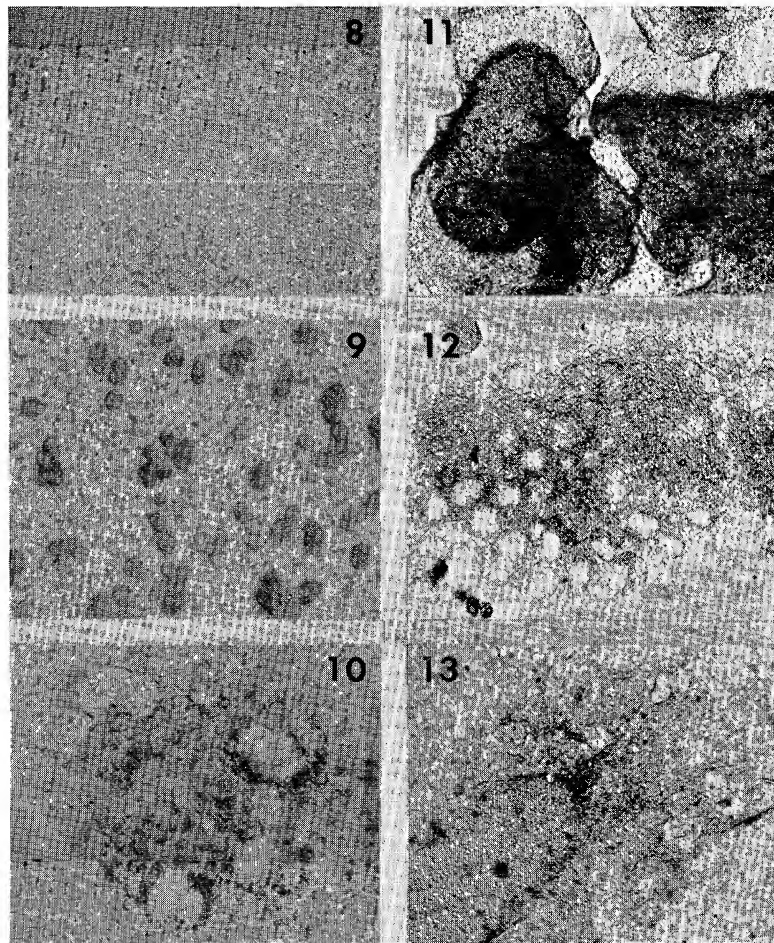


FIG. 8. Cell suspension prepared by trypsinization of metanephrogenic mesenchyme.

FIG. 9. Same suspension after 18 hours. Reaggregation has occurred.

FIG. 10. Reagggregates 3 days after recombination with dorsal spinal cord. Tubules have formed.

FIG. 11. Massive culture of metanephrogenic mesenchyme and dorsal spinal cord, separated from each other by a millipore filter. Mesenchyme (dark) on top of filter, spinal cord (light) below filter.

FIG. 12. Same culture after 48 hours. The spinal cord has been removed to permit visualization of the induced tubules.

FIG. 13. Induction of tubule formation by dorsal spinal cord in reagggregates obtained after disaggregation of tubular mesenchyme shown in Fig. 12.

conversation" to describe the communication system of differentiating cells, a term which seems an ideal choice, since it portrays a picture without defining the medium used to create it. This indeed is the state of current knowledge in this area. It is our hope that as we look more and more at the picture we will learn how it might have been made.

Before embarking further on our consideration of embryonic communication, let us briefly discuss some of the major methods by which communication is achieved in biological systems. Most obvious to most of us is the nervous system type of control, basically a cell-cell communications network of exceeding complexity. A second method is that exemplified by hormonal mechanisms, involving essentially the diffusion of substances which are then transported by a neutral carrier and have a selective effect on certain responding systems. The third method, not as well known or defined, is that of cell and tissue interactions such as we see in embryonic systems. Many features are common to all three, and the applicability of such terms as "feedback," "activation," "all-or-none effects," "sub-threshold activity," "facilitation," "diminution" and the catch-all term "spontaneity" can be well applied to all three of these. But the fact that there is much resemblance between these methods of communication must not let us be misled into thinking that the mechanisms are identical. Discreteness, sensitivity and limits are different, and there are vast areas where one, and only one, of these systems contains the ingredients necessary to meet the requirements posed by a particular problem.

Now let me introduce the system which is involved in our present work. This is the development of the mouse embryonic metanephric kidney (Fig. 1-7). The morphogenesis of this rudiment has been analyzed in a series of studies by Dr. Clifford Grobstein.⁽³⁾ The mouse 11-day metanephros is made up of two components, the mesenchyme, which will give rise to secretory tubules, and the ureteric bud, which will form collecting ducts. In tissue culture, the intact rudiment develops characteristically (Fig. 1-3). Grobstein has shown that development of secretory tubules by the mesenchyme is dependent on an inductive stimulus exerted by the ureteric bud. Thus, by separating the mesenchyme from the bud (Fig. 4) with the use of protein-digesting enzymes, he could show that the isolated

mesenchyme would not produce tubules (Fig. 5). Re-uniting of the mesenchyme with ureteric bud, however, resulted once again in tubule production. Studies on the specificity of the inductive effect exerted on the mesenchyme showed that inductive activity could be displayed by only a few tissues, and by none of the nonliving materials tested. One of the most active inductive tissues was dorsal spinal cord (Fig. 6, 7), and it is the combination of metanephrogenic mesenchyme and dorsal spinal cord which has been used extensively in the further analysis of the induction and organization properties of this system.

In the course of the characterization by Grobstein of the nature of the inductive interaction between dorsal spinal cord and metanephrogenic mesenchyme, two findings are of direct significance to our further discussion. (1) Induction effects can occur across certain membrane filters in the absence of morphologically demonstrable cell contact, and (2) induction is adequate to ensure persistence (stability) of tubules in the mesenchyme if dorsal spinal cord has been present for 30 hours.

The preceding discussion has dealt with the rise of new properties in the kidney mesenchyme, allowing it to form tubules. It must be emphasized, however, that the described observations were all made on tissues, on groups of cells, rather than on individual cells *per se*. It is by no means justified to assume that changes exerted on tissues by tissues, leading to new tissue stability, are necessarily identical to changes, interactions and stability of the individual cells that comprise these tissues (cf. 6, 7). Grobstein⁽³⁾ has suggested that a minimum amount of mesenchyme must be affected to produce a single tubule. Similarly, unpublished studies demonstrate that a minimum amount of spinal cord well above the single-cell level is necessary to produce a recognizable effect on metanephrogenic mesenchyme. These two findings leave open the question of whether the tissue interactions are merely quantitatively larger counterparts of individual cell activities or are rather special functions of the mass, not directly paralleled in the individual cells.

To test the properties of cells rather than tissues we have recently applied techniques of tissue disaggregation and reaggregation to this system.⁽⁴⁾ Kidney mesenchyme was dissociated into cells (Fig. 8) by the use of trypsin⁽⁵⁾ and allowed to reaggregate overnight

(Fig. 9). Such reaggregated mesenchyme was then combined with dorsal spinal cord in standard fashion. Tubules were induced in such reaggregated mesenchymal masses (Fig. 10). Since, after dissociation, cells were able to reorganize into an inductively functional mass, it was concluded that the ability to respond to the inductive stimulus exerted by dorsal spinal cord was a cell-stable property. It must be emphasized, however, that the experiment was not able to distinguish between interactions between cells and interactions between masses of cells since reaggregation is rapid.

Previous studies of Moscona⁽⁵⁾ had suggested that kidney tubules, at a later stage of development, are cell-stable, since disaggregation followed by reaggregation leads to immediate reorganization of these cells into tubular structures. This suggests that while metanephrogenic mesenchyme is stable at the cell level in terms of reorganizing into a homogenous mass of cells capable of responding to inductive stimuli, the end product of the tubule differentiation is a group of cells which after dissociation can reorganize not into a homogenous mass of cells but into an organized, tubule configuration. Induction is seen to produce a change in the inherent self-organizing properties of the mesenchymal cells.

Knowing that cell-stability existed at two stages of the process of tubule formation—the pre-induction stage and the differentiated stage—it raised the possibility of analyzing the stability of newly arising properties. Knowing that tubules are stable at the tissue level after 30 hours of inductive interaction we decided to test whether the new tubules were, at this time, stable at the cell level also. Preliminary results of experiments along these lines are highly suggestive. Kidney mesenchyme was allowed to form tubules by trans-filter interaction with dorsal spinal cord (Fig. 11). After 48 hours, the mesenchyme was removed. At this time well-formed tubules were present (Fig. 12). A few tubules were isolated at this time, and continued to remain as tubules throughout the remainder of the experiment. The rest of the tubular mesenchyme was disaggregated and allowed to reaggregate. One half of the material was isolated; it grew as a homogenous, non-differentiating mass. The other half was allowed to re-react with dorsal spinal cord. In due time tubules appeared in the mesenchyme (Fig. 13). Thus it can be suggested that the properties of tubule stability were

resident in the tissue mass, but were not yet stable at the cellular level.

Grobstein has presented the idea that stability may proceed from group stability to intrinsic cell stability.⁽⁶⁾ He writes: "The alternatives of group and intrinsic stability, of course, are not exclusive. The first may be more important in some instances, the second in others. Or, what seems quite likely, stabilization may begin by group mechanisms but be increasingly supplemented by intrinsic stability of individual cells." Our results are consistent with this suggestion.

While a variety of possible mechanisms for group effects can be postulated, there is increasing implication that extra-cellular materials may be involved. The viewpoint that extra-cellular materials are mediators of inductive interactions has been expressed by Grobstein and is supported by his characterization of the transmission properties in the kidney tubule inducing system.⁽³⁾ The suggestion has been made⁽⁴⁾ that induction leads to a change in the aggregation behavior of cells. The results presented here lend further support to the idea expressed elsewhere⁽⁴⁾ that the same materials involved in induction may be involved in the control of cellular organization and reorganization.

Let us finally once again approach the question of communication. The group (tissue) as evidenced by minimum inductive mass and reactive mass, seems to be the unit of initial embryonic communication. Transfer of information (induction) first leads to a group change as seen by the demonstration of group stability in the absence of intrinsic cell stability. The change from group control to unit control may well represent a necessary step in the process by which the embryonic mechanisms become replaced by more adult mechanisms of control and communication.

SELECTED REFERENCES

Documentation is clearly incomplete. References to earlier work can be obtained from bibliographies of cited publications.

1. M. ABERCROMBIE, in *Chemical Basis of Development* p. 318 (ed. by W. D. MCELROY and B. GLASS) (1958).
2. P. WEISS, *J. Embryol. Exp. Morphol.* **1**, 181 (1953); *Intern. Rev. Cyt.* **7**, 391 (1958).
3. C. GROBSTEIN, *J. Exp. Zool.* **130**, 319 (1955); *Exptl. Cell. Res.* **13**, 575 (1957).
4. R. AUERBACH and C. GROBSTEIN, *Exptl. Cell. Res.* **15**, 384 (1958).

5. A. MOSCONA, *Proc. Natl. Acad. Sci., U.S.* **43**, 184 (1957); *Scientific American* **200**, 132 (1959).
6. C. GROBSTEIN, in *The Cell*, p. 437 (ed. by J. BRACHET and A. E. MIRSKY) (1959).
7. J. P. TRINKAUS, *Am. Naturalist* **90**, 273 (1956).

DISCUSSION

BERG: Since this is an interdisciplinary meeting, I am very interested in knowing if, by changing the culture, you affect the kind of reorganization that you get? In other words, can your environment affect the sort of pattern that you obtain?

AUERBACH: It does so, terribly. As a matter of fact we take advantage of the fact that the culture setup is probably not the best one. The fact is that what we get in the system is most easily characterized and worked with if, say, mesenchyme does or does not form tubules. After reimplantation of mesenchyme into another animal, our reading of a response is complicated by a whole range of unknown and uncontrollable entities. Tissue culture is not the best method for getting the best kind of differentiation; but it is a consistent one. We assume that we characterize, at least in part, the type of phenomenon taking place in normal developmental processes.

As far as changing the type of reorganization pattern of individual cells reaggregating is concerned, Moscona has published beautiful experiments describing the reaggregation processes.

MANOFF (*Air Force Cambridge Research Center, Bedford, Mass.*): Has any chemical been isolated by which these spinal cords are inductive of behavior in human beings?

AUERBACH: This is one of those questions I hate to answer because I have to admit the answer which is this: we have spent a lot of time—unsuccessfully. It is one of those things where at the end you say, "Activity has never been isolated in any form, but we cannot preclude the possibility that——." We have tried all sorts of things. We can't do it. We tried to isolate at different pH, with different kinds of media, and at different temperatures. The thing that happens is this: as soon as the spinal cord cells are dead, there is no demonstrable activity.

CHAIRMAN SCHMIDT: It is true, isn't it, that a wide variety of influences have been tried, meshes, filter layers, and all sorts of things? This is not a poorly investigated field?

AUERBACH: No, by all means. If you want a more complete characterization of what does happen there Grobstein has shown that the activity passes across the filter only if the filter is less than 80μ thick, if the pore size is not restrictive to macromolecules, and only across a gap if the gap includes an area of at least $\frac{1}{10}$ mm diameter. Cellular contact is not seen in terms of the electron microscope. You can keep going on with the properties. The only trouble is, you can't get activity with anything that is non-living, at least not as yet.

FURTHER CONSIDERATION OF CYBERNETIC ASPECTS OF HOMEOSTASIS*

STANFORD GOLDMAN

Department of Electrical Engineering, Syracuse University, Syracuse, N. Y.

THE purpose of this paper is to discuss certain topics in the application of control system theory and information theory to the analysis of homeostasis.† It is assumed that the reader already has a working knowledge of both of these theories and we shall only discuss the ways in which they apply to homeostasis.‡ As a by-product of this investigation, certain new points-of-view in control system theory will come to light.

1. PROPORTIONAL, DERIVATIVE AND INTEGRAL CONTROLS IN HOMEOSTASIS

A concept which is not as widely appreciated in biological circles as it should be is that proportional, derivative and integral controls may all be used in homeostasis. This is illustrated in Fig. 1, which is a diagrammatic representation of part of the blood glucose control system of the body. At the right of the *blood glucose* compartment are six emerging information lines, in accordance with this point-of-view. Let us say that the desired (quiescent) value of blood glucose

* This work has been sponsored by the Information Systems Branch of the Office of Naval Research.

† The name "cybernetics" is used to describe the entire field of control and communication (information) theory, whether in the machine or in the animal. The maintenance of a prescribed internal environment in the body of an organism, in spite of wide fluctuations in its activity and in its external environment, is called "homeostasis."

‡ A previous report "Cybernetic Aspects of Homeostasis" by Stanford Goldman, Syracuse University Research Institute, Report No. EE494-581T1 (Jan 1958) includes a survey of the elements of control system theory and information theory for biologists who have no previous acquaintance with these theories. Most of this earlier report is included in a chapter with the same title, in the treatise *Mineral Metabolism*, which is scheduled to be published in the spring of 1960 by Academic Press (New York).

concentration is around 100 mg per 100 ml. When the actual value in the blood glucose compartment is above the quiescent value, this information may be considered as going out on the $\Delta C(+)$ line to those parts of the regulatory system which react to a higher than quiescent value of blood glucose. When the value of blood glucose concentration is low, the analogous information goes out on the $\Delta C(-)$ negative line. The $\Delta C(-)$ line is distinguished from the $\Delta C(+)$ line because its information does not necessarily go to the same set of organs.

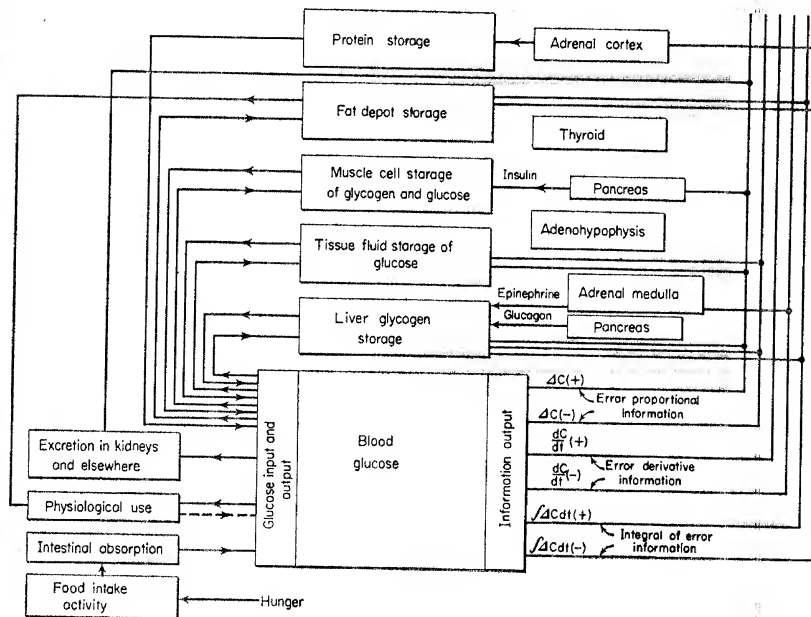


FIG. 1. Part of the glucose control system.

Next consider the error derivative information, dC/dt . When the concentration is falling (dC/dt is negative), this information is transmitted to the adrenal medulla, which then secretes epinephrine that is carried to the liver; the liver in turn converts stored glycogen to glucose, which is released to the blood. It is possible that $dC/dt(+)$ information goes to the pancreas to stimulate the release

of insulin, but whether $\Delta C(+)$ or $dC/dt(+)$ information or *both* stimulate the release of insulin is not clear from the literature.

Integral-or-error information, $\int \Delta C dt(+)$ and $\int \Delta C dt(-)$, both apparently get to the fat depot organ, the former possibly stimulating lipogenesis and the latter stimulating fat mobilization. Negative integral-of-error information gets to the protein storage of the body by way of the adrenal cortex hormones to stimulate gluconeogenesis.

The foregoing paragraphs in conjunction with Fig. 1 have indicated some of the pathways by which information on the blood level of glucose gets to the body reservoirs and other regulatory devices which keep the blood glucose concentration under control. Of special interest is the fact that all six types of information indicated in the figure appear to be used in the regulation of blood glucose; to some extent, at least, the body distinguishes between them and makes different use of them.

There is evidence to indicate that the liver glycogen is acted upon in a different way by proportional, derivative and integral controls. If the liver is perfused by blood relatively free from epinephrine, it appears to be stimulated to store glucose as glycogen if the blood glucose level is high, and to release glucose from glycogen storage, if the blood glucose is low. This is a type of proportional control. On the other hand, the release of glucose from stored glycogen, as induced by way of epinephrine, seems to be a derivative control, i.e. a response to $dC/dt(-)$. Finally, lipogenesis in the liver appears to be an integral control.

In the foregoing paragraphs, we have classified certain known controls of the body glucose as proportional controls, others as derivative controls, and still others as integral controls. Although the statements made seem reasonable, we actually have relatively little quantitative information in the literature upon which to base them. When experiments are made on a regulatory system, it would be desirable from the point-of-view of servo theory to determine whether a control is of the proportional, derivative or integral type. For example, it would be of interest to know how the pancreas responds to the level of blood glucose, to the rate of glucose rise, and to the time integral of excessive glucose concentration. The same information would also be of interest about the adenohipophysis and other regulators. The present author is not competent to judge how much of this information is already in the literature. It would

appear, however, that a proper servo classification of the available details of the control system would help in revealing the nature of the chemical and biological mechanisms involved. Although it may be very difficult to obtain the data required for the above mentioned classification, it may be helpful to keep this classification in mind as an objective.

2. THE USE OF ANTAGONISTS IN HOMEOSTATIC SYSTEMS

In linear electrical feedback systems, a negative error is in all respects similar to a positive error except for a change in sign, which means merely a change in direction of current flow. The same feedback mechanism which corrects a positive error will likewise correct a negative error. This ability to handle positive and negative errors with the same mechanism is not a universal property of feedback control systems. In biological systems, the mechanism to correct positive errors is frequently different from that correcting negative errors. In such a case, one mechanism may be called the antagonist of the other. In biochemical systems, it is relatively easy to add something, but relatively difficult to remove it. It is therefore very convenient to be able to add an antagonist and thus get the equivalent effect of removal of the first controlling agent. The use of a controlling agent plus an antagonist makes it possible to obtain more rapid and accurate control, since the antagonist can be introduced to prevent or cancel overshoot, and in general to balance out residual errors.

In the glucose regulatory system, one unidirectional controlling agent is insulin. Insulin lowers the blood glucose. A hormone of the adenohypophysis, in conjunction with an adrenal cortical hormone, operate as general antagonists for insulin and have the effect of raising the blood glucose. The adrenal cortical hormone apparently promotes gluconeogenesis, thus making more glucose available. The mechanism of action of the adenohypophysal hormone is not certain, but one knows that it tends to cancel the effect of insulin. It is not known whether this adenohypophysal hormone is secreted in response to a low blood glucose, or a falling blood glucose, or to some other cause. Furthermore, it is generally believed that another hormone, glucagon, originating in the pancreas, but in different cells from those secreting insulin, is also important in the

control of blood glucose. Glucagon stimulates the conversion of stored liver glycogen into glucose for release into the blood. Part of the action of the adenohipophysal hormone may be that it stimulates the release of glucagon. The foregoing discussion indicates that while physiological antagonisms exist in the control systems of the body, the antagonists are not necessarily arranged in pairs having opposite functions.

3. HOMEOSTATIC ANALOGUES OF INSTABILITY— THE PATHOLOGICAL DISPLACEMENT OF HOMEOSTATIC CONDITIONS

While the use of antagonists has certain advantages described in the preceding section, it also at times involves the danger of giving rise to instability. The word instability, in this case, has a special

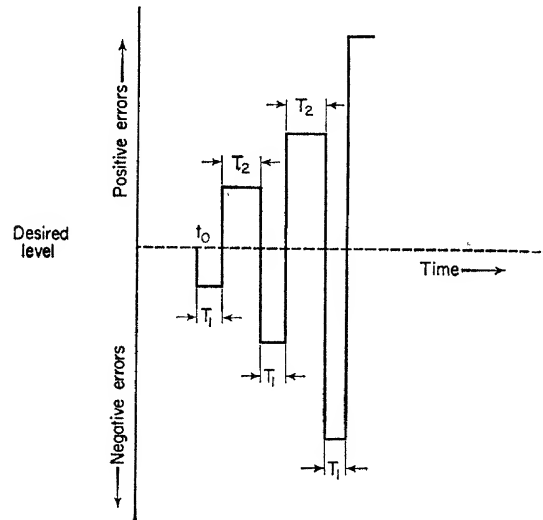


FIG. 2. The growth of error in case of instability.

technical meaning which is more or less equivalent to the terms "vicious cycle" or "runaway process." The following example, depicted in Fig. 2, will serve as an illustration. At time t_0 , a system

characteristic falls below its desired level and a controlling agent starts to act to move the characteristic in the direction to correct the error. Let us suppose that it takes the controlling agent approximately a time T_1 to do its job. For simplicity in discussion, we have idealized the situation in Fig. 2 to show the entire action of the controlling agent as taking place precisely after a time, T_1 , has elapsed. Let us furthermore suppose that the action of the controlling agent is so strong that it overcorrects to cause an even greater error in the opposite direction. At this juncture, the antagonist of the controlling agent comes into play. Let us suppose that the antagonist requires a time, T_2 , to act and let us furthermore suppose that it also is such a strong agent as to overcontrol and cause an even larger error in the opposite direction. After the time, T_2 , the original controlling agent then comes into play again, this time to correct an error larger than the original error.* As the process continues back and forth, the errors continue to grow in magnitude. Ultimately the growth stops because either the controlling agent or the antagonist is no longer able to cope effectively with the very large errors. The final state may be either an oscillation back and forth between large errors of opposite sign, or the system may be locked in a state of large error from which it cannot correct itself, because either the controlling agent or the antagonist cannot correct the final error. This is a story of instability.†

The most dangerous situation from a stability point-of-view arises in the case of a control which is maintained by the balance of two *strong* antagonists. When operating properly, such a control is both fast and accurate. However, the stronger the control (the greater the feedback effect) the greater the danger of instability when the balance between antagonists is impaired. In the case of Fig. 2, it is a timing error which prevents the desired balance. The actions of the controlling agent and its antagonist start too late so that instead of balancing each other, they act regeneratively. Timing errors are probably the most common cause of instability in technical systems, but in homeostatic systems, other causes are probably

* It is not actually necessary that the error should grow in each half oscillation, as shown in Fig. 2, but only that it should grow in each full oscillation in order to illustrate this type of instability.

† This same type of feedback system is actually put to use in the design of electronic oscillators and frequency generators.

more important. Furthermore, in the case of homeostasis, instability is not usually exhibited as an oscillation, but rather as a pathological displacement of the homeostatic balance. Such a situation may occur whenever, because of some fault in the organism, the control device, instead of decreasing the error actually increases it.

Certain types of edema appear to be examples of homeostatic instability, as described above. In these cases of edema, excessive retention of sodium is accompanied by excessive retention of water. The hormone, aldosterone, is believed to cause the kidneys to tend to retain sodium in the body instead of allowing it to be excreted in the kidneys. Greater than normal amounts of aldosterone have been observed in the urine and also in the blood of patients with edema. The overproduction of aldosterone is, however, not the original malfunction involved. It would therefore appear that this overproduction is an error of the regulatory system, in as much as reduction of aldosterone would seem desirable. There is evidence to indicate that the secretion of aldosterone increases in response to a fall in the effective circulating volume of blood. In a healthy man, a decline in the effective circulating volume of blood involves a loss of sodium from the body. An increased level of aldosterone then causes the retention of sodium by the kidneys. This increases body sodium, which in turn tends to increase blood fluid volume and ultimately the effective circulating volume of blood.

A possible explanation of the excess of aldosterone associated with edema would be as follows. First, there is a fall in the effective circulating volume of blood, without a loss of sodium from the body. For example in cardiac edema, the loss in effective circulating volume may first be due to a weakness of the heart pump. In nephrotic edema, the loss in circulating volume may be due to renal loss of blood protein. This loss of protein causes a transfer of fluid from the blood to the tissues. In either of these cases, the regulatory system then responds by increasing the level of aldosterone, which increases the total body sodium above the normal level. Most of this excess sodium does not remain in the blood, but enters the tissue fluid and in turn causes an increased retention of water in the tissues. The original cause of the low circulating volume has, however, not been corrected and any rise in the effective circulating volume obtained by the abnormally high

level of total body sodium still does not bring it up to the desired level. Furthermore, the accumulated tissue fluid probably impedes the circulation of blood. The regulatory system therefore continues to supply excessive aldosterone, increasing blood sodium, tissue fluid sodium and ultimately tissue fluid volume until marked edema has developed. Whether or not the foregoing is the true explanation of various edemas, it illustrates the possibility of a disease of this type being due to a vicious cycle of servo instability.

The most powerful controls (i.e. those having the largest feedback effects) are likely to be those which regulate hormone creation and destruction (or excretion). The reason for this is that the hormones themselves are usually strong controlling agents. Because of the magnitude of the compounded feedback effects thus involved, errors in the systems which bring forth or eliminate hormones are especially likely to give rise to instability.

The foregoing considerations point up the importance of the destruction processes for hormones or other controlling agents. To look at this in a more elementary way, we note that the effect of a controlling agent will be proportional to time, if it works at a given rate. If such a controlling agent is not destroyed (or excreted) or balanced by an antagonist, it may have a cumulative effect which is great enough to cause a pathological displacement of a homeostatic balance.

4. REDUNDANCY IN HOMEOSTASIS

One aspect of information theory which can be recognized in many biological phenomena is the use of redundancy for error reduction. Thus in Fig. 1 we see five different mechanisms for the storage of glucose or its equivalent. Even though each of these has a different specific function, if any one of them gets out of order, the others will work in the direction of correcting the error caused by it. The type of redundancy displayed here, where each redundant element has a different method of operation, is a very effective protection against major errors, since it is unlikely that many of the redundant elements will get out-of-order at the same time or due to the same cause. This type of combination of qualitative and quantitative redundancy represents a safety factor against disease or malfunction.

5. QUANTIZATION AND RELIABILITY IN HOMEOSTATIC CONTROL

A matter which is of considerable interest in our discussion is the difference in properties of quantitative and qualitative information with regard to transmission errors. Quantitative information expresses magnitude. If quantitative information is expressed as the ordinate of a signal, say as a signal voltage, and if the signal is attenuated in transmission, then the information is changed because there is a loss of signal voltage. On the other hand, qualitative information is less susceptible to errors caused by attenuation. Thus a large *W* and a small *w* are both *w*'s. A large apple and a small apple are both apples. Qualitative information is generally less susceptible to various errors arising in information and control transducers, such as noise, backlash, hysteresis and starting errors. Qualitative information is also very versatile in translation. For any or all of these reasons, it is of interest to look for evidence of the translation of quantitative information into qualitative information before transmission in homeostatic control.

Binary information is of the nature of qualitative information in that it is unaffected by attenuation. It is thus immune to errors caused by fluctuations in magnitude. It is also basically very simple. These characteristics would appear to make binary information desirable for use in the body regulatory systems. Thus in Fig. 1 we may expect that the type of binary information which is transmitted is the choice between say,

(*a*) glucose concentration too high,

(*b*) glucose concentration not too high,

and the choice between

(*A*) glucose concentration rising,

(*B*) glucose concentration not rising, and so on.

The presence of some specific chemical in the blood stream or at some organ site may indicate which alternative (*a*) or (*b*) is actually the case. The translation into binary information may well be made by an endocrine organ, or it may be that the endocrine organ itself is already stimulated by binary information. The actual choice between alternatives (*a*) and (*b*) would be made by some threshold operating mechanism. An investigation of the extent to which homeostatic information is binary would certainly be of interest.

When information is translated from one language into another, an equivalence is set up between messages in the message alphabet of one language and those of the other. If each alphabet consists of a discrete set of messages, the equivalence can readily be set up even though the messages are of very different form. Thus XVIII may be translated into 18. However, when information is transmitted in the form of continuous variation (equivalent to continuous curves), it is extremely difficult to translate the information into another language unless the second language is of basically the same form as the first. On the other hand, if a mechanism is available to translate a continuous variation into a quantized representation, further translation into other languages then becomes a much simpler matter.

More than likely, much body control information consists of combinations of signals, such as

- (1) blood glucose is falling, plus
- (2) respiration is rapid, plus
- (3) pulse rate is elevated, plus others.

Such combinations of symbols probably represent important body control information, and more than likely they are made up of binary elements, as they are in the above example. The translation between physiological and psychological forms of such messages would also be simplified if they consisted of groups of binary elements.

6. EXPERIMENTAL QUESTIONS CONCERNING HOMEOSTATIC INFORMATION

In the foregoing sections we have considered certain ways in which cybernetic ideas apply to homeostasis. We shall continue this general objective in the present section, but we shall emphasize the formulation of questions to which it may be hoped experimental answers can be found.

In the first place, the discussion in Section 1 suggests that experiments be done to determine whether any particular control is of a proportional, derivative or integral variety. Alternatively, if it be desired to find out what are the controls of a particular homeostatic system, such as the blood sugar, it would seem appropriate to apply proportional, derivative and integral stimuli separately and see what happens.

If experiments of either of the foregoing types are performed, the question of time scales is important. For example, if the blood sugar is raised in order to study proportional controls, it cannot be raised too rapidly without incurring the probability of stimulating derivative controls, and it cannot be raised too slowly without the possibility of stimulating integral controls. Very likely it is not practical or even possible to separate the various types of controls completely, but the objective may have to be merely to emphasize the effects of one class of controls. The possibility of a fairly complete separation would certainly be greater if the controls are threshold operating devices. Certain preliminary estimates of time scales will have to be made and these must be corrected as the results of the experiments unfold.

Another general class of questions raised by our earlier discussion concerns the translation of control information. Where does the translation take place? What is the form of the translated signal? How is the translation accomplished, i.e. what are the physical and chemical details of the translation process? Suppose, for example, that the primary control information is the existence of a greater than quiescent concentration of blood glucose ($\Delta C+$). In order for this information to be used for control purposes, it must be translated into a form which is suitable for the control system. This translation could conceivably take place in the blood itself, but more likely it takes place in either some monitoring or some utilization site. The translated signal may consist of some specific chemical or it may be an electrical signal in the nervous system. In the case of ($\Delta C+$), it is most likely that a translated signal is insulin secreted into the blood stream by the pancreas, the pancreas acting as the control transducer.* The details of the translating mechanism are not yet known, but they would certainly be of interest. In any complete control loop, several successive translations of the control information may, of course, take place. Needless to say, physiologists did not wait for cybernetic theory in order to realize the importance of the foregoing questions.

* Experiments by E. Anderson and J. A. Long, *Endocrinology* **40**, 92-97 (1947) give strong support to the belief that the pancreas itself is the control transducer between blood sugar level and the secretion of insulin. The question of whether the pancreas responds to ($\Delta C+$) or to $dC/dt(+)$ is, however, not so clearly decided.

Our earlier theoretical discussion suggests that each homeostatic control operation be examined experimentally to see if it is a threshold operating device, and if so, to determine at what threshold level it operates. Furthermore, each control which is a threshold operating device should be examined to see whether it is a purely binary, yes or no, device, or whether it has a greater response for a value of stimulus considerably in excess of the threshold. Care

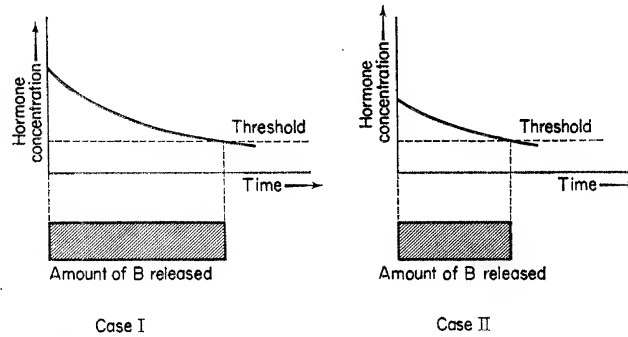


FIG. 3. Proportionality of response illustrated.

must be taken in such an investigation to avoid certain pitfalls of interpretation. For example, suppose that organ A releases substance B at a certain rate into the blood when a particular threshold level of hormone C is exceeded in the blood. Figure 3 shows in this case how a large concentration of hormone C may cause the release of a relatively large amount of substance B; while a smaller concentration of hormone, but still above threshold, will cause the release of a relatively smaller amount of B. The reason for the difference is that it takes a longer time for the elimination processes acting on the larger amount of hormone to bring its concentration below threshold. As a net result, the response in substance B may be more or less proportional to the amount of hormone introduced, even though the hormone is a threshold operating device, having the same effect for all concentrations above threshold and having no effect below threshold.

Another item which cybernetic considerations emphasize for experimental work is the study of the mechanisms of destruction or elimination of chemicals, such as hormones, which act as control

signals. An important related question arises as to whether any of the control mechanisms operate on the rate of destruction of any regulating hormones, such as epinephrine, insulin, and the adrenal cortex hormones. If so, errors in the operation of these control mechanisms could easily give rise to diseases which are homeostatic analogues of instability, because of the large feedback effects involved. Finally, in the study of any disease or symptom, such as edema, hypertension, diabetes, arteriosclerosis or arthritis, especially if the disease is not apparently of infectious origin, it seems worthwhile to examine what part servo instability may play in the picture.

DISCUSSION

CHAIRMAN SCHMIDT: I especially enjoyed hearing you expound on my favorite hobby of redundancy as a control means. But I think the biologists shouldn't let too many of the mathematical and physical scientists get away with a one upmanship trick they have been using very successfully, that of defining the ground rules and making the experiments fit into one or another of them. That is, by saying a thing must be either an integral or derivative or proportional control. This, of course, requires conformity. I think they should thrust before you quantitatively these nasty functions which actually are exhibited in the biological world and require the development of mathematics appropriate to them, from the highly nonlinear, very nasty, double-valued systems.

GOLDMAN: There are a couple of things to be said about that. In the first place when you say something is a proportional control you don't mean necessarily that the response is proportional to the error, but merely that the information to which a proportional control responds is proportional to the magnitude of the error. The magnitude of the action of a proportional control need not be proportional to the magnitude of the error, although that happens to be true in a linear system. In more general cases, such as the one shown in the diagram here, a proportional control merely responds to the magnitude of the error, which in this case would mean that it responds to a change in glucose concentration level.

SCHMIDT: You don't mean either all or none, yes or no? I think you imply monotonic.

GOLDMAN: Yes, in general you do mean monotonic. I am not arguing for the fact that physiological controls are as I have stated here, I merely suggested that these ideas might be of value when you are studying homeostatic controls.

JONES (*Northwestern University*): With reference to slide number one you have shown three possible error signals, proportional, derivative and integral. You have also shown that they could be of either sign. In a number of physiological regulators we have looked at, the presence of both the proportional and derivative signals are regularly found. For instance, about two-thirds of the fibers of the retina produce a derivative type of signal. Derivative signals are found in certain proprioceptive units and in certain thermal receptors. The presence of integral signals, however, is one which I have not met and I wonder if you could throw any light on that?

I might say in passing that the closest approach to this is in the so-called adaptive response of a sensory element following the pictures that Adrian has given showing certain rates of adaptation in nerve pulse frequency and these, if you look at them, come very close to a lead network.

Now a lead network has an integral term in it but it is not a pure integral signal.

GOLDMAN: Well of course the types of control I was talking about are homeostatic controls. I was not talking about muscle control or nerve control in general, although a lot of the same ideas will apply. Now in the cases that I showed there, the type of integral control I talked about was, let's say, the control of the fat depot storage and the control of a gluconeogenesis, that is, the making of glucose from body protein. It is likely, at least, that those are integral controls. If you have a high blood sugar level for a long time, fat will start to be formed because of it. If you have a low blood sugar level for a long time the making of blood sugar from body protein and the use of body fat for metabolic purposes will be stimulated. These effects do not start immediately, but they probably become stronger, the longer the high or low blood sugar level condition lasts. It is therefore likely that here we have a response to the time integral of the blood sugar level "error."

Now, by and large, I would say that adaptation would tend to be a type of integral control. But if you are going into real detail, of course, it is not entirely a question of an integral control but generally a question of the entire past history. The past history of a system is the complete story of which the integral is only one of the parts that determine the adaptation.

POWERS (*Veteran's Administration Research Hospital, Chicago, Illinois*): I have an example of the neural integrator. It is well known and has been known for a long time under another name. That is, a reverberating network, which has been postulated for use as a short-term memory and so forth, sometimes even for a long term. But if you feed a constant frequency of neural impulses into a so-called reverberating network and then examine the frequency of one of the neurons of this net, you will find it increases with time.

Now, this isn't perfect, or the ideal integrator, but at least it is a lagging network.

JONES: Maybe I misinterpreted your figure. I thought that the signals you were showing were sensory signals depending for their pick-up on the state of blood glucose and my particular comment is on sensory pick-ups. One finds the derivative and the proportional, but I do not know of integral signals.

GOLDMAN: Well, my diagram by and large did not consider sensory pick-ups particularly. For example, if you allow blood to go through the pancreas with a high level of glucose you will get the secretion of insulin without any neural intermediary at all and many similar things occur without a neural intermediary. Some controls, however, do have a neural intermediary. But I wasn't specifying whether that was the case or not.

DAVIS (*Syracuse University, Syracuse, New York*): It might be worth making a comment, due to Weiner, that the distinction between the response to an instantaneous value versus an integral really often lies in our hypothesis. When you try to make this mathematical judgement it is in your hypothesis, for instance—a thermometer doesn't really measure the temperature now. If you suddenly change the temperature for a microsecond, it doesn't record that.

FEEDBACK THROUGH THE ENVIRONMENT AS AN ANALOG OF BRAIN FUNCTIONING

GEORGE H. BISHOP*

*Washington University School of Medicine,
St. Louis, Missouri*

SYNOPSIS

IN THE following essay several proposals will be made in addition to the implications of the title.

(1) The physiologist cannot deal with subjective response directly, but there are reasons for inferring that the physiological patterning corresponding to subjective response is similar to that involved in the organization of motor behavior.

(2) Something is known about the simpler brains and about the phylogenetic origins of higher brains. Successive structures and functions have been added to the more primitive apparatus.

(3) The many different elements of activity acquired in this evolutionary development have such diverse properties and such complicated interactions as to indicate that no very fixed or specific paths of activity are available for stereotyped behavior to specific stimuli. Rather the result of stimulation seems to be a statistical average of the activities of many variable units, the final sum being perhaps more specific than the activity of any one unit.

(4) The possibility of self-regulation may be related to this non-specificity of response, permitting a number of possible responses among which choice can be made.

(5) The final selection and organization of the response pattern most appropriate to a given environmental stimulus needs constant monitoring from the environment through the sense organs, especially to the changes in the environmental pattern of stimulation resulting from activity initially attempted. This feedback through the environment is proposed as an

* This work was supported in part by a grant from the Supreme Council, 33rd Scottish Rite, Northern Masonic Jurisdiction, U.S.A., through the National Association for Mental Health; and in part by contract between Washington University and the Office of Naval Research.

analog of a corresponding feedback to memory as a record of past environmental stimulation.

(6) The organizing of such a response pattern involves time, during which an initial and tentative choice is modulated and elaborated as the resultant activity is being carried out. It is suggested that this proposition can be generalized to be applicable to subjective activity.

A. BRAIN AND MIND

IN CONSIDERING the brain as a mechanism which thinks, from the physical or neurophysiological view, we may be guided by our knowledge of some of the other functions of the brain. First, from considerations of comparative physiology we can infer that most animals don't think, at any very high level at least, but all of them *act* quite expertly. Their brains are obviously organs devoted chiefly to the control of motor behavior and to the organization of useful patterns suitable to the environment of the moment. Second, the brains of the higher animals differ from those of the lower chiefly by an increase in number of cells and probably by greater complexity of their interconnections. The fundamental properties of neurons, their structure, and many of their patterns of activity are strictly comparable. Finally, the structural differences between the different areas of a given higher brain amount to relatively minor differences of proportion and arrangement of elements. In particular, this applies to those areas more recently expanded in the primates and man.

One could not infer from the structural relations what contributions these areas make to brain function. The evidence as to this function is derived from the effects of their ablation, which results in alterations of behavior patterns. Removal of one area of brain only interferes with consciousness without abolishing it. Removal of these recently acquired association areas alters the emphasis or choice of behavior patterns without grossly preventing their execution. In a sense, such removal reduces or simplifies brain functioning as if the decrease in amount of tissue available resulted in a decrease of complexity of pattern and flexibility of choice, with some degree of specificity assignable to different areas. For all such reasons, any one of which may be equivocal or inadequate, the best inference is that the degree of consciousness possible, the competence of the brain in generating subjective behavior, is a function primarily of numbers of cells and complexity of their interconnections.

Since the basic and primary function of the nervous system is the organization of patterns of movement (and other efferent activity), we may proceed on the inference that organization of motor patterns is the prototype of the organization corresponding to subjective thinking. Both occur, if not in the same neurons, at least in similar complexes and patterns.

How then does the brain determine, select and organize the patterns of physiological activity? What determines its choice of one pattern or another? We should first know what a brain is, where it came from, what it is made of, and what kinds of activity go on in its parts. Finally we should want a wiring diagram. We might then be able to recognize a pattern of activity in nerve cells that corresponded to the making of a decision. I can answer none of these questions in detail, though some information can be presented that may point to the answers.

Much subjective experience is useful as evidence of physiological activity. When we perceive a sensation, we can relate it to a pattern of physiological activity in a chain of neurons from the periphery to the cortex. When we see how the subject responds, we can define the activity of another chain of neurons extending from the cortex to the periphery. What goes on between these chains, in addition to the mere passing along of nerve impulses, should correspond to the making of a decision or choice, which may or may not be a conscious one.

Here we are in a dilemma, for the items of subjective activity and of physiological activity are presented in different terms, and even in different categories, from which are drawn many nontransferable implications. The simplest element of subjective experience must correspond to the activity of a large number of neurons, but we do not always know which neurons nor what pattern of activity they are employed in.

The only neural activities which we can at present analyze accurately are those simple responses following arbitrary stimulation, or those that are so elementary that they could not be identified uniquely with any but the most elementary subjective events. This will be apparent presently when certain reaction patterns of which cells are capable will be described.

I have made some attempts to persuade psychologists of my acquaintance to attempt a breakdown of the subjective activities

they deal with into still simpler elements, in the hope that they might arrive at functions simple enough to correspond to the most complex activity which the physiologist can accurately study. We could then draw more reliably on subjective experience as evidence of neuronal patterns of activity. So far the attempts have not been very successful, and for the time being we will have to infer from the simple patterns we can deal with what more complex patterns might correspond to subjective bits of behavior. We are only a little better off in dealing with *objective* behavior above the reflex level.

B. WHAT IS A BRAIN?

When the unsophisticated person speaks of the brain he thinks of a mass of nervous tissue above the neck. When the anatomist speaks of the brain he thinks principally of the central mass above the brainstem, consisting of the thalamus, basal ganglia, and cortex. A psychologist's picture of the brain tends to emphasize the cortex as the presumptive substrate of consciousness and of memory. Animals had brains before they had any of these structures except a knob of thalamus at the end of what is still called their brainstem. A brain in general is a structure primarily devoted to the analysis of afferent information, and to the integration of this information into patterns of useful behavior. All the brains, and there have been a series of them, have originated in relation to the special sensory systems. They still function basically as analyzers of sensory information and integrators of motor patterns.

C. THE EVOLUTION OF BRAINS

The nervous system of the vertebrate is a tubular structure with bulges along it. The lowest bulge above the spinal cord is the medulla and above this are arranged the cerebellum, the tectum of the brainstem, the thalamus, the basal ganglia, and cortex, in serial order. These enlargements were developed serially from posterior to anterior, each (excepting the basal ganglia) originating from one or more sensory nuclei (Fig. 1). Each sensory "brain" as it developed became increasingly dominant over the structures below it. Each performed two main functions; it elaborated and analyzed the details of the sensory information contributed from its own

level, and it acquired connections from other sensory levels. It thus became, to some extent, a sensory analyzing and integrating apparatus for the body as a whole, not for its level alone. On the outgoing side it activated and directed more or less of the whole motor apparatus into patterns of action appropriate to the situation.

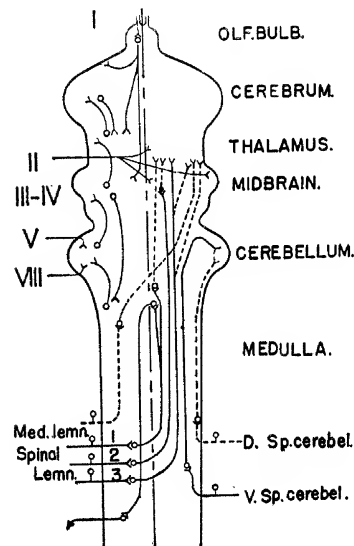


FIG. 1.* Diagram of the vertebrate nervous system indicating the main divisions and certain of the afferent tracts. Roman numerals stand for the sensory nerves from whose central nuclei the levels indicated developed in the course of evolution. 1-3, the chief sensory branches of the spinal lemniscus, present in premammals. The paths indicated by dashed lines have been developed only in mammals. Similar paths of visual and olfactory sensory systems and certain connections between levels are indicated on the left half of the diagram.

Thus any brain, however primitive and elementary, makes decisions and choices among alternatives and integrates motions into movements. It may or may not do this consciously, and its choice may or may not be predetermined by a specific pattern of stimulation.

The last brain developed at any stage of evolution tends to become the dominant one, and uses the previously developed centers below it as agents, each, however, still contributing its characteristic capability to organization of the whole behavior pattern. Thus the

* Figures 1 and 2 are reproduced from Bishop, Relation of nerve fiber size to sensory modality. *Journ. Nerv. Ment. Dis.* 1959, 128: 94. Figure 3 is reproduced from Bishop, chapter 20 in *Reticular Formation of the Brain*, 1958, Ford Hospital Symposium. Little Brown & Co., Boston. Figure 5 is reproduced from Clare and Bishop, Dendritic circuits. *Amer. Journ. Psychiat.* 1955, 111: 826.

medulla finally specializes in the control of visceral organs, the cerebellum in control of the muscular system and its smoothness of operation. The tectum adds the contribution of the visual system to the pattern of motor activity. The thalamus develops beyond a visual sensory center acquiring, in addition to the optic pathway, afferent tracts from the rest of the periphery of the body (Fig. 2). Finally, the cortex, developed first as a facilitatory adjunct to the olfactory sensory system, acquires secondarily profuse

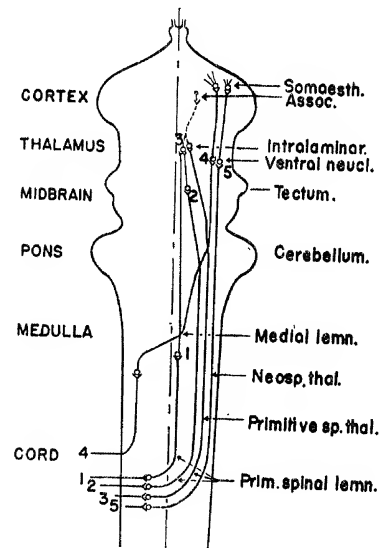


FIG. 2. Similar to Fig. 1, but details of five parallel sensory paths making connections from sense organs to higher centers are included. Only two of these, 4 and 5, project sensations directly to cortex. The others, representing paths inherited from our ancestors, terminate in the thalamus or below it. They are apparently relayed to cortex, but secondarily, and by unknown connections.

connections from and to the thalamus. The latter, starting as a visual nucleus, becomes a co-ordinating center for all the sensory systems of the body, by relays from the lower centers and from the afferent paths already established at lower levels. The cerebral cortex proper is unique in some respects, one of which is that it receives little or no information *directly* from the periphery. The

primitive sensory connection to the cortex is the olfactory path to the piriform lobe, from which the general or somatic cortex developed secondarily. The activities of the other afferent systems are filtered through the thalamus to reach the general or nonolfactory cortex, with more or less alteration of pattern in the process. In fact, so intricately are cortex and thalamus of the mammal tied together by paths in both directions between them, that they may be inferred to function effectively as one organ. It is still possible, under this inference, to think of each of these structures contributing its characteristic aspect of function to a final result in behavior.

Since the thalamus has lost most of its motor connections, while the cortex has acquired a variety of outgoing channels, we may picture the thalamus as the major input channel to cortex (Fig. 2) and see in the cortex an apparatus for further analysis of the information delivered to the thalamus. This is followed serially in cortical functional organization by an integrating apparatus and a motor outflow. As cortical accessories, we must add a memory storage depository and a reservoir of hereditary patterns predisposing the brain to certain preferred patterns of activity. Many of the latter are built into the basal ganglia and operate both by way of cortical connections and also by way of motor outflow from these nuclei into the old brainstem motor system. The more recently established cortical motor outflow, the pyramidal tract, operates at all levels on this older motor system. These levels include the basal ganglia themselves, the brainstem motor levels and the spinal cord. The cortex acts on the periphery only through these levels, rather than on the final effector organs of muscle, gland, etc.

D. THE EVOLUTION OF CORTEX

The cortex, as noted above, originated as an offshoot of the olfactory sensory apparatus, and served first as a modifier or facilitator of olfactory function. This original cortex persists as the direct precursor of the piriform cortex of higher animals. At a stage of evolution somewhere below the reptiles, one region of this primitive cortex established connections with the thalamus, already at this stage a sensory integrating center with a motor outflow for the control of lower levels. This "general" cortex progressively lost most of its original connections to the olfactory system as it gained similar relations to the thalamus (Fig. 3, I). As the general

cortex expanded in the higher vertebrates, it acquired, on the one hand, more specific relations to the thalamic sensory centers, and on the other, progressively took over from the thalamus the control of motor activity. To accomplish the latter it must have become something more than a modifier of thalamic activity, probably its original relation. It became progressively a higher receiving station for sensory paths, and developed progressively a motor outflow of its own.

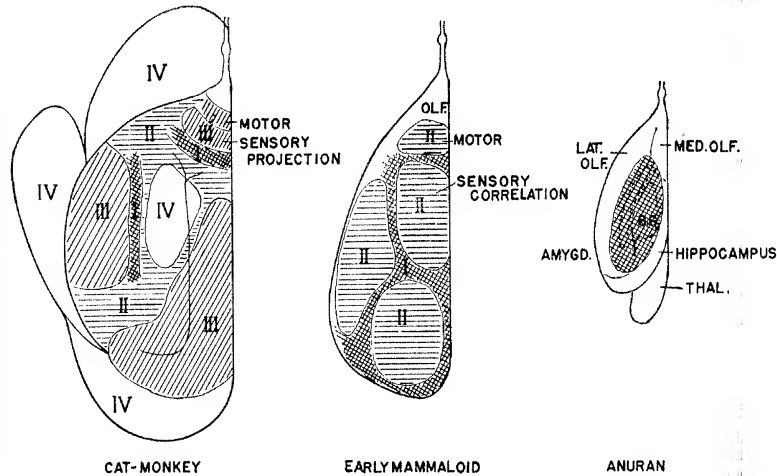


FIG. 3. Stages in the evolution of the cortex. See text.

The first step in this shift of dominance from thalamus to cortex consisted of the specialization in the primitive cortex of three areas. These areas established particular reference to the three great sensory systems of visual, auditory and somaesthetic reception (Fig. 3, II) as already represented in the thalamus. This stage is approximately that of the present reptilian brain, and presumably of the missing links between reptilian and mammalian forms. In the mammal, the next stage (Fig. 3, III) appears fully completed, showing specific projection areas to which the three sensory systems send information through the thalamus by direct relays. Here the sensory periphery is projected in some detail as a map of the body surface reference, spread across the cortex in the case of somaesthetic sense. Corresponding detailed projection of visual field and auditory frequency,

etc. occur in their respective areas. At this stage the cortex has apparently added an improved analyzer of sensory information to what the thalamus had accomplished, establishing its dominance over the thalamus as a sensory apparatus, but still depending on the thalamus for its input.

Not all the sensory modalities are thus directly projected to these specific projection areas. These other modalities, including pain and temperature senses, still seem to be projected to thalamus (Fig. 2). From here to cortex, the more diffuse relation of thalamic and cortical sensory function found in the pre-mammals may still persist. The cortex receives most directly only certain sensory paths for fine discrimination of touch, vibration sense, muscle tension, etc. and corresponding components of visual and auditory stimulation. Their common denominator in the three projection areas appears to be a content of information about the relations of the body to objects in surrounding space, as contrasted to information as to the state of the body itself, exteroceptive information as contrasted to interoceptive. To what extent and in what manner the modalities of sense not represented in the three projective areas are represented in other regions of cortex, we do not know at present.

The latest stage, if not the last in the evolution of the brain, consists in the expansion of the association areas into frontal, parietal, temporal and occipital lobes (Fig. 3, IV). This occurs progressively in the primates and has its maximal development in man. The question is, of what previous structures are these the expansions? They seem to have no immediate relation to sensory perception. Minor differences in structure give no evidence of the character of their contribution to behavior. Results of their ablation appear chiefly as modifications of attitude, interest, and other characteristics so far defined chiefly in subjective terms. Such removals do not result in aphasia as removals of older association areas may, nor do they interfere with ability to execute motor patterns. They are not primarily memory reservoirs. They certainly contribute to the making of decisions or choice, but perhaps as biasing or modulating mechanisms rather than as the primary substrates for integration of behavior patterns. Without drawing sharp lines as to the anatomical or functional limits of these areas, they obviously supply large numbers of neurons available as the substrate of complexity of brain function. They might be

diagrammed as eddies of the main stream of input-to-output flow of activity. As such they would receive stimuli from the rest of the brain, give the activity this generates a characteristic emphasis, and feed this emphasis or predisposition back into the main stream, but with only secondary relation to sensory input or motor output by way of the more basic association areas. They might stand in the same relation to the general integrating mechanisms of the brain as does the most primitive piriform cortex to the olfactory sensory receptor system.

Looking at the overall picture of how the higher brain structures have developed, a block diagram may be presented indicating the possible relations of its main constituents in terms of overall function. For the sake of simplicity of presentation, we may limit this analysis to thalamus and cortex primarily. Starting with a sensory integrating center in the thalamus the thalamo-cortex has passed through several stages (Fig. 4). First occurred the addition of cortex as a modifier to a previously developed thalamic integrating center (Fig. 4-1). Then followed a differentiation of regions of sensory reference in cortex related to afferent receptor centers in thalamus (Fig. 4-2). From this stage on, thalamus and cortex have progressed in parallel, but with dominance shifting from thalamus to cortex, signalized by loss of thalamic motor tracts and gain of cortical pyramidal tract outflow. Both thalamus and cortex differentiate further to take advantage of the acquisition by mammals of a new sensory projective pathway from the dorsal columns of the cord (Fig. 4-3). A relay of this path in the thalamus, and a terminus in cortex itself, again suggests cortical dominance. The development of memory storage, assignable chiefly to cortex, with the resultant capacity for learned behavior, and the expansion of "association" cortex (Fig. 4-3A) related to the integrative function of the system rather than to sensory analysis, adds a group of complex modifiers to the substrate of the basic response patterns. A group of hereditary or instinctive biases persists throughout, exemplified by, though not confined to, the basal ganglia. Nothing appears to have been lost in this process of development; the process has involved adding more complex organs with resultant greater flexibility and possibility of variable response.

This potentiality of modifiable behavior connotes the increased possibility of choice and decision making. Can one infer, from

such a course of development and the increase of complexity of organization resulting, how decisions are arrived at, that is, how choices are made?

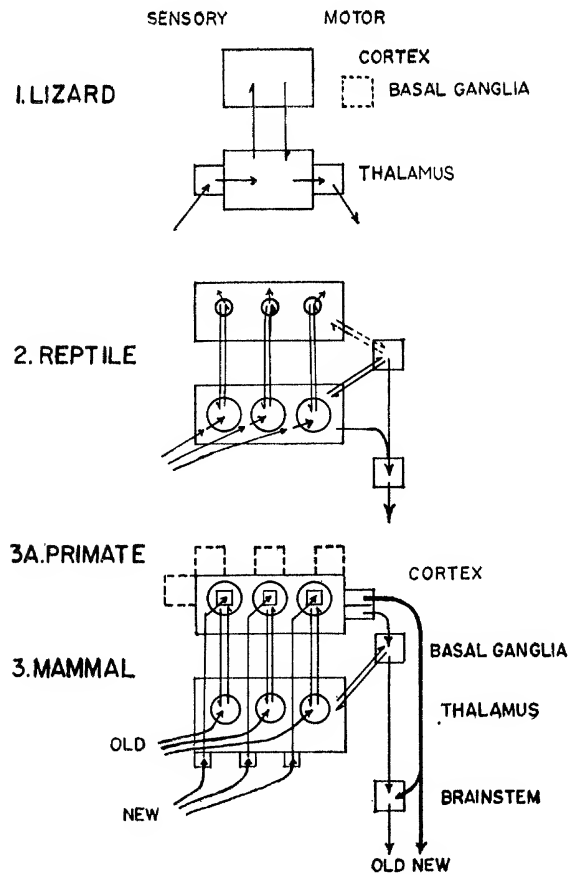


FIG. 4. Block diagram indicating the successive stages in development of, and interconnections between thalamus and cortex. Those labelled "old" are the primitive premammalian paths, the "new" are only developed in mammals. Dotted squares in 3A represent the expansion in the primates and man of the association areas. Sensory on left, motor on right side.

E. THE CONSTITUENT PROCESSES OF NERVOUS TISSUE ACTIVITY

Before one tries to analyze brain function as a whole, some consideration should be given to the behavior of its parts. The unit of the nervous system is the nerve cell or neuron, and each neuron consists of four main divisions, each with different properties (Fig. 5). The nerve fiber or axon extends from the cell body to an effector organ, which may be another neuron, a muscle, or gland cell, etc. Its sole function is to conduct an all-or-none impulse, that is, one which is of constant size whatever the intensity or character of the stimulus. It carries its messages in code, as patterns of successive impulses variable in frequency and number. In the central nervous system its impulses originate usually at the cell body, which responds to electrical depolarization of the cell membrane. The greater the depolarization, the higher the frequency of response. This cell body depolarization is induced by an entirely different process occurring in other fibers branching from the cell body, termed dendrites. At each point where an axon impinges on dendritic tissue an impulse arriving over any terminal of that axon produces a depolarization locally of the dendritic membrane, usually by means of a chemical secreted by the axon terminal. The cell body itself may receive axon terminals, and thus act as dendritic tissue. These four parts, cell body, axon, axon terminal and dendrite, thus constitute the functional neuron. The cell body in addition to being typically the generator of all-or-none impulses also supplies the metabolic energy for the activities of all these cell structures.

The region of contact of an axon with a dendrite is the synapse. The chemical depolarization induced there is not all-or-none, and two or more successive impulses at short intervals can sum their effects to a stronger depolarization. Repetitive impulses can in fact maintain a continuous depolarization at a given synapse. Such a depolarization at a synaptic point draws current from the cell body, tending to depolarize this passively. One dendritic system of one cell may have many synapses on it, and their depolarizing effects are summated at the cell body. If enough active synapses draw sufficient current a threshold depolarization is induced at which level the cell body is excited to the production of an all-or-none impulse, which is conducted into its axon. Stronger

and longer-lasting depolarization generates a train of repetitive impulses. Many different fibers from different regions of the brain may terminate on any one neuron, and their effects are then summated at the point of generation of impulses. Other synapses may be inhibitory, acting to reduce current flow from the impulse generator, and thereby to reduce or prevent its depolarization.

The progress of a message through a series of neurons thus involves coded sequences of all-or-none axonal impulses alternating with non-all-or-none (graded) steady or variable depolarizations of dendritic tissue. The pattern of the message transmitted over each cell is chiefly determined in its dendritic phase, for here the messages from many neurons interact and are summated to determine the impulse pattern of the next axon. It is further probable that no neuron in the central nervous system ever acts independently of other neurons, and its level of excitation is so complexly determined as to be in effect continuously variable, as contrasted to the all-or-none character of the axonal impulse.

The axon can be compared to a vacuum tube employed to register only off-on values as in a digital counter, while the synaptic activity of dendrites has more the properties of a tube working as an analog device. Even in parallel arrangements where a group of fibers activates a group of cells in the nearest approximation to a one-to-one relation, connections of branching fiber terminals permit the impulse in each fiber to contribute to the dendritic depolarization of a number of cells, so that no such arrangement as a single chain of neurons exists in the working nervous system. Moreover, the afferent fibers are in general of different sizes and presumably deliver correspondingly different complements of energy to any given cell. The net result is that messages carried over a system of cells having a common function must be capable of an indefinite series of values, and the end result must have, even for simple activities, some of the attributes of a statistical summation rather than of a specifically transmitted pattern.

In some neurons at least, perhaps in most, more than one group of axons impinge, each group on a different region of the cell with different results. The typical pyramid cell of the projection areas of cortex has a major apical dendrite ramifying at the surface of the cortex and a ring of basal dendrites about its cell body (Fig. 5). Different groups of fibers activate these two regions. Stimulation in

certain regions of thalamus and cortex results in activation of apical dendrites only. The latencies of response indicate that two groups of axons, both small but in two size ranges, lead to the dendritic terminals at a given locus of cortex. These responses differ in several characteristics, indicating that the synapses the different fibers make with dendrites are of different character. Fibers from cortex to thalamus have similar connection with thalamic neurons.

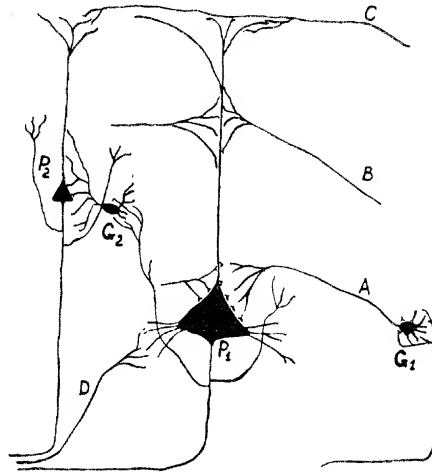


FIG. 5. Diagram of cortical pyramid cells P_1 and P_2 , and interconnecting cells G_1 and G_2 . A to D, different paths that may send impulses to different regions of any one cell. The fiber rising from P_1 to the surface of the cortex is the apical dendrite, that passing downward from the cell body is the axon or nerve fiber, with collateral branches arborizing within the cortex. The main axon passes through the white matter below cortex to activate other regions of the nervous system.

Still a third group of axons, constituting the sensory path relayed from the periphery to cortex, activates the basal dendrites or the cortical cell bodies themselves. When these different systems of fibers are activated separately and in different time relations it is found that the larger fibers impinging on the lower cell body and its dendrites tend to produce impulses there in a one-to-one relation, while the smaller fibers reaching the apical dendrites at the cortical surface typically fail to cause the cell to discharge impulses at all. However, if the small fiber path is activated just before the large

fiber path, the latter causes more cells to respond, indicating that the former path has partially excited the impulse-generating locus making the cells easier to fire by way of the large-fiber afferents. The basal cell connection thus tends to serve as a relatively simple relay, but one whose action is subject to modulation by way of various small-fiber paths to apical dendrites.

In experiments under any conditions short of deep anesthesia, a persistent and nonpredictable fluctuation of the response to constant stimulation of the afferent tract alone may be assigned to the continuous but highly variable summation of impulses from other centers acting on the apical dendritic structure. Such modulation is presumably a major factor in the control and shaping of the patterns of activity organized by central structures, and much of the functioning of the nervous mechanisms may be performed by synaptic activity below the level of overt impulse firing. This dendritic system then acts as a mixer apparatus by which many influences may be imposed on a given process, to appear overtly only in the final resulting pattern.

The above details of cell and circuit functioning have been presented as illustrations of the manner in which patterns of activity, however arbitrarily or randomly initiated, may be altered, modulated in intensity, or even redirected in their transmission through the nervous system. The items cited by no means exhaust the possibilities, though we know little about the details of normal functioning in terms of these elements. There are axons in the mammalian brain of diameters ranging from 0.0003 mm to 0.015 mm, a ratio of nearly 50 : 1, with cell sizes in proportion. There are many different forms of neurons with different distributions of dendritic terminals and of paths ending on them. There are at least four types of synaptic processes distinguishable by their differential depression by different chemical agents, and there are probably many more than we know about. There are circulating hormones that alter neuron excitability. The variety of patterns in which neurons are anatomically related is still being explored, and their number is practically legion. Even these patterns are probably not functionally fixed ones, and there is good evidence that one neuron may be connected into different functional groupings depending on the pathway over which excitation is delivered. There are many preferred channels whose activities seem specific

when an animal is under anesthesia or otherwise reduced in capacity to respond, but which are no longer isolated from interference from other brain activity in the intact animal.

The more one attempts to analyze the patterns of response characteristic of what might appear to be specific cell groupings, the less it appears that there are such arrangements as will give any very specific patterns as a fixed component of behavior. Rather the relations of neurons to each other involve extreme flexibility of participation in multiple patterns. Activity flows through the more complicated neural networks with somewhat the fluidity of waves over water. Only at the periphery do we find more unique connections and more inflexibly fixed paths. Certain more or less fixed inherent or acquired barriers or predispositions result in preferred but not obligatory courses for the flow of activity through this complicated network. Otherwise these courses exhibit a considerable latitude within which they are easily shifted about by impulses so multiple and so various as to approximate randomness. That final patterns of behavior so purposeful and so predictable can be obtained from such an apparatus may imply that this random activity is scattered about certain means, and the statistical distribution of variable elements of functional activity is more specific than are the units of activity themselves.

F. THE FLOW OF ACTIVITY THROUGH THE BRAIN

Many activities are initiated externally to the central nervous system, in bodily organs or in the sensory periphery. Any activity so initiated, or however initiated in fact, will enter the nervous system main channels by some specific route, and thus be predisposed to take some preferred general direction through the higher centers. This action-initiating stimulus will arrive against a background of sensory or other activity which may be below threshold for overt motor response, but still capable of predisposing the integrative mechanism to a corresponding set or bias. The central result in any case will be a significant modulation of the input pattern. The stream of activity, so modified, must then flow in part into the reservoirs of stored past experience (memory), and the resultant pattern, modified by matching of the present stimulus with previously registered patterns, will be reflected back into the main stream. A somewhat similar influence or series of

them may be assigned to other modifiers of pattern such as are inferred to reside in the various expansions of the "association" areas, the frontal, temporal, etc., auxiliary regions of the brain. The pattern so reorganized must further run the gauntlet of the hereditary or built-in directive or patterning apparatus exemplified

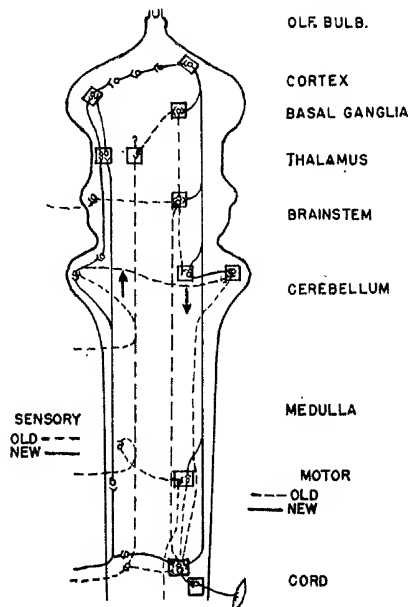


FIG. 6. Diagram indicating the connections at various levels between sensory and motor systems. Ascending sensory and other afferent paths enter from left side of diagram, descending motor outflow to muscle etc. leave on right. Squares indicate some of the principal centers of the nervous system, particularly on the motor side. The full line on the right represents the mammalian pyramidal or corticospinal tract, which sends fibers from the cortex to all the lower centers of the old pre-mammalian motor system, still functional in mammals.

in connections of thalamus and cortex with the basal ganglia. The resultant stream emerges from the motor area of cortex, to be projected on the periphery.

We have outlined here, however sketchily, only that branch of the afferent stream arriving at the thalamocortical level of the

nervous system. However, the cortical outflow over the pyramidal path is not the only determinant of motor pattern. An older motor system passing downward from basal ganglia to spinal cord is accessible at various levels to activation by branches of the ascending afferent paths (Fig. 6). The cortical pyramidal system acts on this older system at various levels also and overt motor activity is the resultant of these two streams of stimulation. Since the higher decision-making levels are what concern us here, the execution of the directive, however arrived at, is of secondary concern. Without the normal activity of the rest of the motor apparatus, the cortical motor pattern would be incomplete as an organizer of motor behavior. This is obvious from the effects of lesions in this older motor system whose derangement can effectively disorganize, and even paralyze bodily movement. In other words, the motor pattern transmitted from the cortex is alone far from adequate as the transmitter of purposeful directives, but acts rather as a modulator working through lower motor levels, each contributing its characteristic component to the whole pattern.

The course of transmission of activity through the higher centers has been treated above as if it consisted of a single sequence of activity occupying serially one after another locus. Instead, all central activity involves time, and no single volley of impulses transmitted through the nervous system can result in precise and purposeful behavior. In particular it takes time to make decisions. A constant or persisting stream of impulses passes through the central apparatus, modifiable at any stage as it continues to flow. The first overt motor action results in modification of the stream of afferent impulses through a change in the relation of the body to the environment, and this modified activity alters cortical activity. Thus the process of pattern formation continues as the activity proceeds. This is in effect a feedback, through the environment, of the initial and continuing results of central activity, and its effect is to monitor the activity toward the goal originally set for it. In so doing, it may also modify the goal itself, by correcting or amplifying it or even reversing the decision originally arrived at. In a nervous system so essentially unsuitable for the making of precise and adequate responses to the limited information usually available to it, such a means of continuous correction and adjustment is an appropriate and necessary adjunct.

One may observe the action of this feedback apparatus in any number of motor activities. A simple one consists of a test applied by clinicians for the detection of malfunction of the cerebellum. The subject sits at ease with his arm extended, and is asked to touch his forefinger to the tip of his nose. The normal subject does this smoothly and with considerable precision. He is then asked to do it repetitively, with the intent of making precisely the same contact each time. The subject then closes his eyes and repeats the movement. The result will be a considerable loss in accuracy, even in the normal subject, and greater error if the cerebellum fails to function normally.

When one examines his own behavior in performing this test, several points of interest become apparent. First, the movement itself involves complex and accurately balanced contraction of a large number of muscles in shoulder, arm and hand, and demands the maintenance of this balance, though a constantly shifting one, as the direction and speed of movement changes during its execution. Second, the competence of the movement is measured by the fine discrimination of position. As a final measure of precision of control, each contact sensation is compared with those preceding it. Other sensations from muscles, joints, etc., doubtless contribute to accuracy of movement. One other sensory control, made evident by the loss of precision on closing the eyes, epitomizes the whole procedure. Apparently the eyes are watching the movement of the finger, and constantly induce shifts of pattern of muscular contraction *during the movement itself*, in anticipation of the final contact. When one notices how poor a visual image one obtains of the tip of one's nose by looking down it, and considers the speed with which adjustment must be made to be effective, the fact that addition of visual monitoring of such a movement is effective emphasizes two points. The first is that without this final monitoring even simple acts are inexpertly guided by the brain. The second is that the monitoring is continuous, which implies continual development or adjustment of central pattern during its expression at the periphery. In a sense, the definitive and final central directive is not completely formulated until the resultant motion itself is completed.

The reason for thus examining the motor end effect of central activity is that it is accessible to examination as the central activity

is not, and this peripheral process may serve as an analog of what happens in the brain in making even such decisions as may not be expressed in overt acts, as in "thinking." When central activity with whatever initial slant or predisposition is initiated, the process of further defining the final result, the choice between possibilities, may be affected by a similar feedback within the brain itself. This is saying no more than that when we receive a stimulus to act, what we decide to do, or what we decide to be appropriate, is defined not only by the nature of the stimulus but by the modulation of the reaction to it, subjective as well as objective, with reference to past as well as to present experience, and that this occurs as a continuous process as decisions are being formulated.

G. THE ORGANIZATION OF CENTRAL PATTERNS OF ACTIVITY

If this analogy with the regulation of motor activity during its progress is to be useful, something in the central process of decision-making itself should correspond to the continuous feedback from motor activity to sensory input. The obvious central analogy of

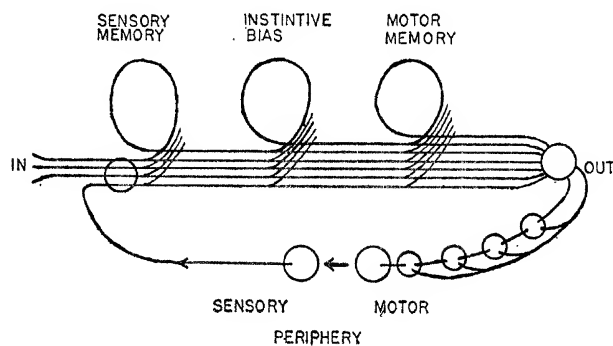


FIG. 7. Diagram indicating the passage of messages through the cortex, modified at successive loci in passage, in the process of determining the appropriate response to a sensory stimulus. See text.

current input information resulting from current activity is similar past experience registered in memory (Fig. 7). As a stream of activity, however initiated, flows across the nervous system, its

course must be capable of redirection by comparison of present stimulus with past experience. Similarly, the *anticipation* of the result of central activity so determined must be compared again with the records of past final results of previous similar activity. In this respect, a memory reservoir serves the same function as the environment in the case of motor output discussed above. Into this reservoir a partially formulated decision sends a warning of its possible or probable results, to be further limited or modulated in accordance with the memory of previous results of similar decisions. It is of course incidental whether this memory is a conscious process or not; as here used, it implies only stored information accessible to brain functioning.

We may arrange some of the things the brain does, in simplest terms, in order of decreasing knowledge of how the brain does them. We can then attempt to extrapolate from the better to the lesser known.

(1) Stimulation at two widely different input points gives widely different motor activity. Presumably channels from input to output are then maximally separated anatomically and functionally, and the "choice" of response is chiefly based on locus of stimulation. Difference of responses is maximally assignable to difference of initial bias, resulting from spatial relations.

(2) Choice between responses to stimuli of different modality at the same locus may be made on the basis of intensities of the stimuli, or of fusion of the two resultant sensations to a new complex. Each stream of activity may then be pictured as modulating the other, over a wide and continuous range of possible results.

(3) Redirection of activity in response to the sensory result secondary to motor activity; feedback through the periphery.

(4) Choice between two or more possibilities to a given stimulus determined by learning, memory of past experience. This must involve modification of path by memory of the *result* of previous responses, as well as by memory of previous stimuli to response; a central feedback through memory as a "peripheral" environment; peripheral, that is, to the path of activity initiated by the stimulus. Past experience stored in memory may then function similarly to current sensations fed back from environmental changes due to motor activity. It must similarly contribute to the further organization of a pattern of activity even during its formulation.

The proposal here is that the final pattern is not an integral event occurring as a single sweep of impulses across the brain, but is a structure which grows in time by successive contributions, from various areas of brain or from various functional organizations of neurons. Each successive contribution is first induced by the flow of impulses eddying from a main stream into a reservoir of stored information or of inherited predispositions. The effect of each such contribution is fed back into the main stream to modify the course and content of the stream. This then determines the courses of successive eddies. Whether or not overt movement results, whether the activity is only manifest as subjective thinking or feeling, whether in fact the process reaches consciousness, is incidental to the proposal that some such process as this is probably the way the brain functions.

H. SUMMARY STATEMENT

In this account it has been emphasized that complexity of organization in nervous tissue should correlate with flexibility of behavior. Another way of putting this is that the more ways there are in which a complex structure can operate, the less fixed and stereotyped its operations are liable to be. In the control of overt movement the central nervous system does not operate precisely or invariably. It requires monitoring from the periphery through the sensory effects of the movements being executed. Memory is notoriously inexact, learning is progressive and often inaccurate. Subjective activity is prone to hallucinations and misconceptions. To obtain reproducible results from experimental stimulation an animal is anesthetized or decerebrated. On the other hand, without the lower level functions of the nervous system the higher levels are incapable of adequately organizing behavior.

One inference that might be drawn from such complications is that the possibility of self-regulation, that is, the possibility of organizing its own behavior, on the basis of its own content and structure, is related to this relative lack of fixed patterns of reproducible activity. The typical motor activity of an animal is a reaction to the environment, through the sense organs and sensory pathways. The adequacy of that activity is checked and controlled by the changes produced in the relation to the environment, detected again through the sensory pathways. The higher animals have

progressively developed and increased the capacity to remember past experience, and this stored information seems to substitute for, combine with, or exert an influence comparable to that of the sensory messages from the environment. With increase of ability to store past experience, and corresponding increase of choices of action available on this basis, the process of developing useful patterns must become more liable to variation, and presumably to error. To compensate for this, a more highly discriminating apparatus for selection of components of a final pattern must be required, a more elaborate mechanism for reference to the items suitable to serve as guides to present action. Even at best, with increase in flexibility of operation and in numbers of possible alternatives, the final result must tend to be less specific and less predetermined by a specific input.

We do not know by what process the brain selects or reacts to the appropriate fraction of all things remembered although a logical sequence of propositions can be made. First, the more nerve cells available and the more complex their interconnections, the more elaborate and comprehensive is the potential substrate of memory. Second, the more available this expanded substrate becomes to a given stream of activity, the greater the uncertainty of selection, and the greater the freedom of choice resulting. Third, under these circumstances a means of further checking on the appropriateness of the resultant activity should be and is available in the sensory feedback from motor activity. We may infer that a corresponding reference of a provisional choice of action to a pertinent region of memory storage could be the central counterpart of this sensory monitoring.

DISCUSSION

STECK (*Sandia Corporation, Albuquerque, New Mexico*): I wonder whether you can comment on the present state of knowledge concerning the mechanism by which information is stored—the storage of information in the brain and even subsequently recall?

BISHOP: I can only quote Dr. Karl Lashley, who said he spent his life looking for the engram and the only thing he could tell you is where it was not. I haven't improved on that one bit.

CHAIRMAN SCHMIDT: Would you care to elaborate on this to say whether you think it is in specific places or whether it is totally distributed?

BISHOP: This is what puzzles me. I can only tell you I am troubled about it. Memory cannot be the function of one neuron. There are just too many things against that. The engram can't be a mark on one neuron.

Now, how does memory, especially conscious memory, how does anything conscious, involve two or more neurons which are separate units and which express their relation only by working on something else? What is it they work on? It puts consciousness, and memory, right outside the physiological nervous system.

If we follow logically what we know about physiology we just say we don't know anything about this. However, we all assume physical or chemical traces in nervous tissue corresponding to memory.

SCHMIDT: Further questions?

WEYL: Does anyone know what the brain does as reported subjectively by those who experience it when deprived of all sensory inputs?

BISHOP: What does a brain do when it is deprived of all sensory inputs? It goes crazy more or less. It apparently doesn't know what to do and proceeds to operate without the usual sensory controls, and soon wanders irrationally. Like any physiological system which is accustomed to being stimulated, the brain presumably becomes hypersensitive when normal stimuli are cut off. Like the visceral system, for instance, which is principally activated by the autonomic nervous system. When this is cut away, the viscera become hypersensitive to circulating adrenalin, and still function actively, if erratically. I would say the hallucinations, presumably erupting spontaneously from memories of past experience, are abnormal because they lack the guidance of current experience. This might be taken as an example of the dependence of the brain on constant checking of its current activity by reference to immediate experience. The feedback through the periphery that I discussed, of the results of voluntary activity, is one variety of such checking of central activity against current input. I would propose this as a general principle of central activity. Without this, the central nervous system is not a self-regulating one.

Now if you cut the afferents to the brain, you are doing the same thing to the brain as cutting the sympathetic nerves to the viscera. You are letting it go without activity until finally, I suppose, it tends to get into random activity and there you get into hallucinations and the sort of things that are reported when one is deprived of all sensory input, as far as this is possible.

Now, if you cut a cat's brain stem, cut all the afferents you can in that cat, you still leave a fairly competent cat who will still eat and walk around and so on somewhat like a normal cat, but that cat will have a pattern of activity in the brain that is called "spindling." The electrocorticogram pattern, instead of being the alpha rhythm or a modification of it, comes in bursts. First a spindle of oscillations occurs which blocks out the normal pattern, then a stage of quiescence, then another spindle. This is also produced under deep nembutal anesthesia. Anything that stops the brain from functioning due to external sources results in this so-called spindle pattern, or even in what is further called sleep pattern.

In other words, there are signs that the brain's activity, when all nerve inputs are removed, goes into some automatic spontaneous oscillation which doesn't necessarily have any meaning to it. What they mean to the cat, I don't know. This activity is not normal, though it resembles that of normal sleep, and perhaps dreams and activity similar to the hallucinations following deprivation of normal sensations.

SCHMIDT: You had ample opportunity to speculate about these central nervous mechanisms and I know you have both the background in engineering and neurophysiology to consider it adequately. Can you attempt any generalization regarding the properties of these heavily re-entrant negative and positive

feedback systems with respect to their patterns of behavior? This is the thing most engineering designs shy away completely from: a system of just more than a few loops. Here we have clear evidence, I think, of many thousands of short loops, long loops, positive, negative loops, many of them nonlinear, all of them tightly connected. I suppose we are going to have to start making some generalizations about them and I wonder whether you are any better off in this than we were ten years ago?

BISHOP: I don't think that I am. I would say that if any computer designer wanted to take advantage of the brain as a model he would have an awful job on his hands making anything like a brain do what he wants a computer to do, if I understand you right.

The brain is not an accurate organ. Its memory is poor, its recognition is poor, its sensory input doesn't work very well, its motor activity isn't very well organized until one becomes skilled. This is not a good computer. What it does is to respond variably over a range. There is an advantage in the circumstance that it doesn't always do the same thing to the same stimulus. If it misses once in a while it may learn something about how to operate more usefully.

If you ever got a man so well educated, so learned and so highly skilled that he always gave the right answer to every stimulus that happened, never made a mistake, well, I think he would lose his advantage over a computer.

The ability to make mistakes connotes the ability to make choices over a range not accurately defined, and to find new ways of doing things, and that is probably one of the virtues of the imperfections of the nervous system.

CRITCHLOW (*IBM Research, San Jose, California*): I am very interested in these many levels of control. It is interesting to talk about the four or five levels of control you have in the human brain apparently, because computers apparently are built the same way. They start out with several computers and a compiler on top of that and now there are supervisory computers on top of that.

Would you care to comment on the reasons for this evolution if you have any thoughts on it?

BISHOP: I don't know the reason. You must ask God for that. All I know is that this is the way the brain developed. The history of the nervous system tells us something about how the structures came about, but not why they came about that way. I suppose our ancestors, the lizards, needed certain brains for certain things and they got them because of natural selection and evolution and so on and the next higher animal came along and something changed that to a more effective animal by putting something on top of the lizard's brain. We never lose anything. There is nothing in the brain that you cannot find the beginnings of in the larval form just hatched from the egg of the lizard. It contains all the fundamental structures of the human brain. Not the details, but the main divisions and levels.

This thing has been built on, adding one thing to another, and its present status comes out of its history.

To go back of that, you can say what it does, but you can't say why it is there. You can't say the brain added a given part because it needed it; rather that it made use of such additions as it was able to acquire in the process of evolution and natural selection, such additions as had survival value.

FIRST DAY

GENERAL DISCUSSION

Questions Directed to the Panel at Large

CHAIRMAN SCHMIDT: Would anyone suggest some criteria to distinguish self-organizing systems?

Dr. von Foerster looks very anxious to answer that.

VON FOERSTER: I am very anxious to answer that? (Laughter). It depends of course on which way you would like to ask that question. Maybe we should have a little discussion.

SCHMIDT: It is directed to anyone who wants to answer it.

VON FOERSTER: I abstain for the moment.

SCHMIDT: No, go ahead. You answer it first and then we will get some corrections. There will be corrections no matter who answers the question.

VON FOERSTER: I think I pointed out somehow what I meant by self-organizing systems and as I already talked about it I think somebody else should answer it.

SCHMIDT: Well, that is letting you off the hook completely. Does anybody else want to try this first? No; you see you have the floor. Let's hear the comments.

ROSENBLATT: I think Dr. von Foerster defined a province for himself and I think by default he should answer this question.

VON FOERSTER: We must go back to some of the ideas which we already discussed this afternoon. For instance, Dr. Auerbach's paper, this extraordinarily interesting paper on the cells which reorganize themselves after they have been taken apart. I would not have used my funny magnets if I had heard Auerbach's paper first. You see, with the magnets, before I was saying what they really do, I had already given my trick away: that they are magnetized. Assume for the moment you don't know that they are magnetized and you shake them around and they form beautiful lattices and do fantastic things. You will wonder and say, "For heaven's sake, what goes on? It is a very peculiar system." Then comes someone and tells you they are just magnetized in a tricky fashion and you say, "This is a cheater. This is sleight-of-hand."

Now in the case of the cells, we may not know what the principle is that really brings these things about and we say obviously to what we see in front of our eyes: that is self-organization!

The ultimate problem is probably a theory of the structure of the universe referring to the particular structures of the elementary particles of which it is built. Although we are constantly going down the energy drain, we are forming systems of higher complexity.

SCHMIDT: Thank you. Dr. Farley, do you care to comment any further on this?

FARLEY: In that first paper Clark and I wrote anticipating questions of this sort, we defined what we meant by self-organizing systems, and in a very simple

way by merely isolating a system, giving it some sort of test input and then scoring the output in some arbitrary way and then allowing it to undergo some specified experiences and then testing it again and seeing if this score has improved. And if the score has improved we say the system is self-organizing.

Now it is clear that this is relative to the particular test and the particular scores that you use, but I don't see any other way of defining it. Similarly, of course, if it doesn't improve, it is not self-organizing in this respect.

VON FOERSTER: What about learning?

FARLEY: All I can say is it is a model of learning. This is exactly analogous of course to somebody who is taking a course in French. You may give him a test in French before he begins and score that, and after he has been in the course for a month you give him another test and score that.

VON FOERSTER: I would say the whole idea of self-organizing systems is not at all mystical. A self-organizing system is just like a salt solution; if you dry out the water you find crystals are forming and this is, in a sense, a self-organizing system if you wish; but probably this is not the kind of thing you have in mind.

ROGERS: Following Dr. Bishop's paper, Dr. Weyl asked the question, "What happens to a brain which you detach from its environment?" and the question of detaching the brain was answered by Dr. Bishop.

I would like to add some additional comments to this. Fortunately, we don't have to detach the brain from its environment. Another way of asking the question is, what happens when you isolate the entire organism from its environment?

Now back in 1955 and 1956, if my memory of the newspaper articles which appeared at the time suffices, the National Institute of Health ran a series of experiments called "The Dead Man's Float," in which, in effect, a human being was completely isolated from any kind of stimulus whatsoever.

The experiments were run in studies of psychotic personalities and if one takes the position that the psychotic is the person that is afraid of all stimuli that he finds in his environment, perhaps psychotics can be treated if we isolate them from the environment.

Well what was done briefly was that a large tank was built and filled with water which was kept at a constant temperature, and in this tank the subject was immersed. At first the subjects were the designers of the experiment. The person was kept in a rubber suit which completely enclosed him and floated face down with oxygen supplied through a tube.

The experience was reported by the subjects as delicious at first and most subjects fell asleep between 20 min and a few hours after having been put into the tank. Upon waking from sleep, however, the subjects found themselves completely disoriented in space, floating around in this tank they had no way of knowing what direction was up or down. They of course had no visual or auditory stimuli and the skin had adapted to the temperature and pressure of the water to such an extent they could feel nothing as far as moving their arms, trying to get some feel from the water.

The interesting part for this present discussion is what happens to the person's thinking. And the frightening aspect, the frightening result reported by the subjects was that after a while—and some of these experiments went on with the subject in the tank for as long as 12 hours—after a while the subject found it impossible to change the idea he had. One thought would let's say, take hold, and in Dr. von Foerster's terms, with no noise from the environment he found it impossible to change the idea. It just kept going round and round.

Under the circumstances it was easy, of course, to inject material—there were earphone arrangements so that the people watching the experiment could

inject material which was immediately grasped by the subject and that also went round and round. When the subjects were finally released from the experiment it took approximately 24 hours for them to get back to normal in the sense that they could take information which was presented to them from the people outside the experiment and make what we would call rational evaluations of it. That is, compare it with other things they find in the environment, compare it with their own memory.

The access to memory, which one might expect the subject to be able to have even under those circumstances, was totally lacking, and this ties in: I hope it presents another answer to the question you asked and perhaps ties in with some of the comments of Dr. von Foerster this morning.

HANDRIX (*Naval Ordnance Test Station and U.C.L.A.*): Have there been any theories of consciousness on a mechanistic basis? We can conceive of building a machine that could simulate the entire behavior of an organism, if not a man, then something less complex, say a rat. But such a simulator would not, I think, be conscious. Can consciousness be mechanized? Are there any theories or opinions relating to this? Please don't ask me to define consciousness.

ROSENBLATT: I might cite a few things from the literature. Unfortunately, I must ask the questioner to define consciousness or I don't think the question is a particularly meaningful one as nobody has really reached an agreement on this point.

Let me indicate however that there are a number of schools of thought among psychologists as well as automata theorists, as to whether consciousness exists, as to what it is if it does exist, and as to whether we have any reasonable scientific interest in it in any case.

Culbertson, for one, tackles the question very directly and assumes that consciousness does exist and deals with this in terms of a tree of relata spreading indefinitely into the past. I am not too sympathetic with this approach, but I would certainly refer you to this reference which you will find in his book which is called *Consciousness and Behavior*.

The people who have approached this question at all, and I think most people have become increasingly cautious about tackling it, seem to be divided between those who regard it as essentially a private phenomenon, a sort of epi-phenomenon which really has nothing to do with the way the system functions, and those who feel that in some sense consciousness represents a phenomenon which is quite essential to the kind of activity with which it is associated, and that in some sense conscious processes can not only be distinguished from nonconscious processes, but actually perform in different ways and have different roles in the behavior of the organism; that consciousness is in some sense functional and cannot be disregarded as the other group feels it can.

One of my colleagues has suggested that consciousness is whatever people are referring to when they say they are conscious, and I am personally inclined to go along with this. This does not necessarily imply a position on the utility of consciousness. I also feel that when people say they are conscious they are acting in a way which can be distinguished in some fashion from other ways and this probably has something to do with their awareness of the flow of information through their own sensory channels and through their own central nervous system. By this I mean that somebody who can distinguish between a state in which he is not receiving any information at all because he happens to be blind at the time, and a state in which he is not receiving any information because nothing is there, is in a sense being conscious, and that this particular form of consciousness (if we accept this as consciousness) is indeed functional.

It is quite important to know whether sight has failed because of a visual malfunction or whether sight has failed because the eyes are closed or something of this sort. Consciousness then involves presumably a discrimination and consequently information, in the true information theory sense, about the state of the system. If there are no such states which can be discriminated, if there is no information contained in the statement that I am conscious, then the question itself becomes meaningless. We must then in approaching this problem at all, I think, begin by asking, "What are the states about which we are supplying information when we say we are conscious or we are not conscious?" And here again I think we come back to the question of usage. What is it that people refer to about their behavior or about their internal state or about their sensory apparatus or anything else when they say they are being conscious or are not being conscious or have been conscious or have not been conscious of some event?

SCHMIDT: I can't resist the temptation with this array of expert talent here to see if someone here can't postulate some workable hypothesis upon which we can base the interesting 24-hour or one-month periodicities which we find in any biological organized systems and apparently haven't been able to localize biochemically, biophysically or neurologically and yet find prevalent in almost all animals and many plants?

GOLDMAN: There is an experimental connection between the periodicities and some of the hard radiation that gets into the earth.

SCHMIDT: Are you quoting Frank Brown in that connection?

GOLDMAN: I didn't read this, but I listened to a lecture where a man described how he had slices of liver and watched their metabolism and found that it varied in accordance with such periodicities. They finally found they were able to correlate the periodicities with certain hard gamma radiation or cosmic rays coming to the earth at the time. I don't know anything more about this thing.

SCHMIDT: I would say this was less than fully verified.

AUERBACH: I might just mention one thing along these lines. I think this periodicity is fascinating. Any outside source which doesn't take the responding system into account is going to be in trouble. As an example, we have in our department a Dr. Rawson who has been working on this with inbred mouse strains—it is a funny thing, in the dark, you know these cycles are not 24 hours, there is a 23 something hours or 22 something, depending on the strain used. One can set one's watch by the way these animals work in the dark and I don't think any outside source is sufficient to account for the fine strain differences in rhythm. It is a very complicated system and we simply don't understand the mechanism as yet.

SCHMIDT: I think this is an interesting problem related to these organized systems in that apparently many of these properties are essentially independent of temperature and most biochemistry is strongly dependent on temperature and it would be interesting to find systems that don't have this temperature correlation.

VON FOERSTER: My impression is that whenever we deal with very accurate clocks, some strong feedback action is involved. This means that two systems, *A*, *B*, strongly interact, whereby *A* controls *B*, and *B*, in turn, controls *A*. Consider for the moment a simple enzyme system where the activity of the enzyme is inhibited by the products of its activity. This system usually follows the so-called "logistic" differential equation of growth:

$$\frac{dn}{dt} = K n (1 - n) - \lambda n,$$

where K and λ are the build-up and decay constants respectively. Clearly, in equilibrium $dn/dt = 0$ and we have

$$n_{\infty} = 1 - \frac{\lambda}{K}.$$

But the two coefficients K , λ are temperature dependent, usually following Arrhenius' equation:

$$K = K_0 \exp\left(-\frac{E_1}{kT}\right)$$

$$\lambda = \lambda_0 \exp\left(-\frac{E_2}{kT}\right)$$

where E_1 and E_2 are the energies of activation for these enzyme reactions, k and T are Boltzmann's constant and absolute temperature respectively. Inserting these into our expression for the enzyme equilibrium density, we have

$$n_{\infty} = 1 - \frac{\lambda_0}{K_0} \exp\left(\frac{E_1 - E_2}{kT}\right)$$

As you can easily see, if the activation energies are of about the same value, n_{∞} becomes quite independent of temperature, because the exponent in the e -function is so small that temperature fluctuations will show up only as a small correction:

$$n_{\infty} = 1 - \frac{\lambda_0}{K_0} \left(1 + \frac{\Delta E}{kT}\right)$$

SCHMIDT: It is my impression that these carry exponential characteristics with negative and positive coefficients and do not cancel out over 20 or 30°. Isn't that right?

VON FOERSTER: It depends on the activation constants. I think you can get the temperature independence over 50 or 60°.

WEINBERG (I.B.M.): In biological systems it seems that a major obstacle to study is the inability to determine what are the stimuli. Some of the problems we seem to be having with computer models arise from our inability to determine what is the response of the system. We stand in great danger of predetermining our results by the way we select criteria for response. Just as in a biological system, the most significant responses may not be detected by the experimenter. How do we know when a system is responding or what are responses?

FARLEY: I don't believe it is different from any other scientific problem.

ROSENBLATT: It seems to me in most cases we have defined what responses we are interested in when we set up the experiment in the first place. If we don't know the variable which we are concerned with and are interested in investigating there really isn't much point in performing the experiment except as a purely phenomenological investigation. If the experiment is to be more than that then we have presumably defined some measure we are concerned with, and changes in this particular measure constitute the response.

AUERBACH: Don't you think it is of importance to recognize the fact that there are probably many other responses either which we are not interested in enough or we are not aware of that are nevertheless important?

ROSENBLATT: In the case of complex systems that may be highly important, but as Bel just observed, I don't see that this is very different in our own particular

field of discussion, from any other scientific field; there may always be many other variables that may or may not be relevant to the problem.

A VOICE: I think part of the question is there is an enormous literature of physics and physiology and psychology which regards perception as response and not a stimulus. There has been considerable discussion here of putting percepts through a machine. Now the perception is a response, I would think, and I think this is probably the question we have shifting back and forth between responses and stimulus.

ROSENBLATT: But I don't recall anyone has spoken to-day of percepts being presented to the system as inputs. We have spoken of percepts as responses of the system. I realize that it has been traditional in the course of the last few psychological association meetings to have at least one symposium on the nature of the response. Personally, I have found these rather fruitless. It seems to me as I indicated just before, that the response is whatever it is that we are particularly concerned with investigating in a particular case. I don't see that we are necessarily constrained to call one variable the response and another the stimulus or vice versa. It seems to me this is very much relative to the interest of the investigator and the particular problem.

A VARIETY OF INTELLIGENT LEARNING IN A GENERAL PROBLEM SOLVER

A. NEWELL,* J. C. SHAW* and H. A. SIMON†

THE analysis in this paper is part of an exploration of the possibilities for learning and self-organization in a computer program called the General Problem Solver I, or GPS. GPS is a program that incorporates heuristic means for solving a substantial range of problems including, for example, discovering proofs for theorems in logic, proving algebraic and trigonometric identities, and performing formal integration and differentiation. The analysis derives from the following heuristic: To study learning and self organization, take a program that accomplishes a significant task and discover all the ways it can be improved and can improve itself. Heuristic programs are likely candidates for such an investigation, since, by their very nature, they are open to improvement almost everywhere. We might have chosen for study our chess program or LT, our earlier program for proving theorems in logic, but the reason for preferring GPS will become apparent immediately.

The basic learning situation is depicted in Fig. 1. The performance program at the bottom of the figure is GPS. A learning situation requires another program, called the learning program, that operates on the performance program as its object to produce a new performance program better adapted to its task. For GPS, the changes must make it a better problem solver. We will not try to define in general the notion of adaptive change and its measurement. In each particular learning situation we must convince ourselves that the learning program is so structured that it will try to produce better programs, although it may not always succeed, even in the long run.

* The RAND Corporation, 1700 Main Street, Santa Monica, California.

† Carnegie Institute of Technology.

A performance program like GPS is large and complex and is organized from a set of components or aspects. There will be many different learning opportunities corresponding to the different aspects—it being our ultimate goal to discover each of these and

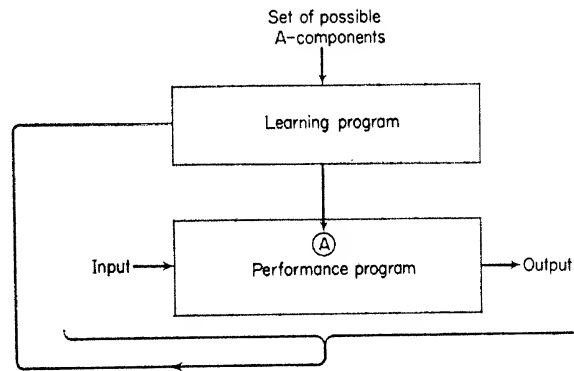


FIG. 1. Basic learning situation.

determine its nature. The learning program for any particular aspect, A , must have access to a set of possible components for the performance program. The set may be given simply by a list, or by the variation of numerical parameters. If the aspect selected for learning is a significant part of the performance system, then the space of possible components will be large and complex. It might consist, for example, of all programs that can be built up from a set of primitive processes. The learning program also must have access to information about the performance program, its inputs, and its outputs. What information is used will vary, of course, but we shall assume that the learning program has essentially complete information about the structure of the performance program and its behavior for a sample of tasks. The learning program may work iteratively over time, selecting candidate A -components, modifying the performance program accordingly, watching the modified program operate, and then repeating the cycle. It need not proceed in this way, however. Although a learning program is constrained, by definition, to produce new and hopefully better programs, it is not constrained to do it any particular way.

Our problem, then, is to construct a program, which we call a learning program, that will make a good selection of an element from the set of *A*'s, so as to yield an effective performance program—in the present instance, a program that can solve problems. If the performance program handles a significant task, if the *A*-component chosen is a significant aspect of the performance program, and if the space of *A*-components is sufficiently rich; construction of the learning program will pose an interesting problem.*

In designing learning programs we are using a particular heuristic:

Generally: that significant learning situations will require learning programs that are heuristic problem-solving programs, in the sense in which that word is currently used in discussing chess and theorem-proving programs.†

More specifically: that because GPS has pretensions of solving a wide array of problems, it may be possible to let GPS be its own learning program, so that the problem of selecting an *A*-component will be a problem of the form GPS can work on.

The purpose of this paper is to follow this heuristic in order to see where it leads. It will become apparent that our analysis is still incomplete, although we have tried to be as definite as we can. Perhaps, even so, we have traced enough of the path so that the reader can evaluate the potentialities and difficulties of this approach.

THE GENERAL PROBLEM SOLVER

GPS and its performance have been described in detail in other publications.^(3,4) We will include here only enough description to make GPS comprehensible to readers who are not already familiar with it.

GPS is a program for working on tasks in an environment consisting of *objects* and *operators*. Symbolic logic is one particular task environment in which GPS can operate. Figure 2 shows an

* Thus the reason some early attempts, like Oettinger's program for conditioned response,⁽¹⁾ have not led very far, although they clearly are learning programs, is that the space of components is too simple and regular. Conversely, some of the interest of Friedberg's learning program⁽²⁾ stems from the fact that the space of components for his program consists of all possible programs—a very large and irregular space.

† Heuristic programs are still best described by example (see Ref. 5, 4).

example of a simple task in this environment. The objects on the left-hand side of the figure are symbolic logic expressions, or propositions; the operators, on the right-hand side, are the rules

OBJECTS	OPERATORS
L1: $S.(\sim P \supset Q)$	
	R1: $A.B \rightarrow B.A$
L2: $(\sim P \supset Q).S$	
	R6: $\sim A \supset B \rightarrow A \vee B$
L3: $(P \vee Q).S$	
	R1: $A \vee B \rightarrow B \vee A$
L4: $(Q \vee P).S$	

FIG. 2. Symbolic logic problem.

that define the admissible transformations of one expression into another. For example, the object, L1, is transformed into the object, L2, by applying the operator, R1. This particular operator is applied by substituting S for A and $(\sim P \supset Q)$ for B in its "input" side, and extracting the corresponding expression (L2) from its "output" side. GPS can solve problems like: "Transform L1 into L4." Figure 2 shows a solution for this problem.

The objects to which GPS is applied need not be logic expressions, nor the operators rules of logic. Figure 3 depicts schematically the general nature of the GPS task environment. Here the objects are geometric shapes, and the arrows show the possible transformations of one shape into another by application of operators. In this environment the problem might be posed of transforming the three shapes on the left end of the figure into the shape at the far right. GPS should be able to operate on any environment where there are "things" that can be transformed or combined into other things by applying identifiable operators or rules, and where the things are *describable*—i.e. have features. The significance of this last

qualification will become evident in a moment, as we explain how GPS operates.

The principal components of GPS are a set of *goal types*: Goal types are used to state problems for GPS and are the major units

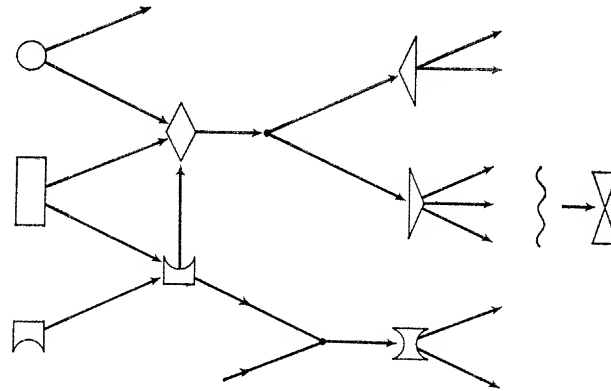


FIG. 3. GPS task environment.

for organizing the problem-solving process. For our present purposes we need to consider only three types of goals: to *transform* one object, *a*, into another object *b*; to apply an operator, *q*, to an

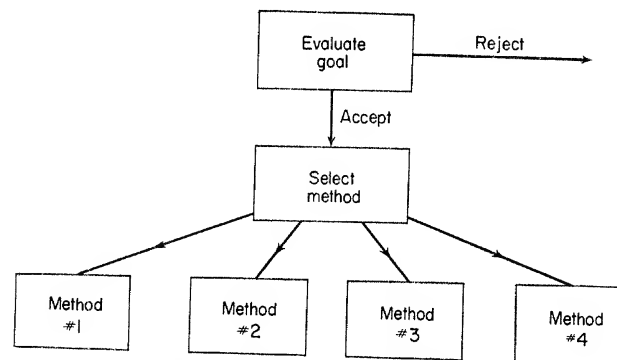


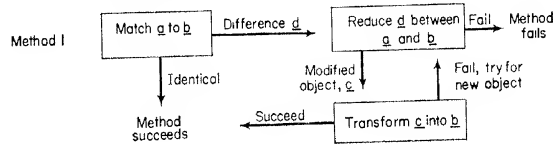
FIG. 4. GPS basic organization.

object *a*; and to reduce a difference, *d*, on an object *a*. At the outset, GPS is given a particular goal (e.g. the *transform* goal in Fig. 2, of changing L1 into L4). It proceeds as follows (Fig. 4):

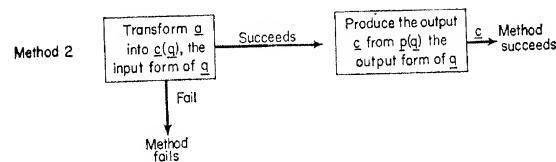
It evaluates the goal to see whether it should be worked on; if it accepts the goal, it selects a method associated with that goal, and applies it; if the method fails, it evaluates whether to continue attempting the goal. If so, it selects another method, and so on.

The main clues to the behavior of GPS lie in the methods themselves. These give GPS a basically recursive structure: methods operate by establishing sub-goals (belonging to one or another of the three goal types) that are (hopefully) easier than the original goal, until a stage is reached where a sub-goal can actually be achieved. When this happens, it represents a step of progress in the goal next above, which can then make progress for the goal above it, and so on.

Goal type No.1: Transform object a into object b



Goal type No.2: Apply operator q to object a



Goal type No.3: Reduce the difference, d , between object a and object b

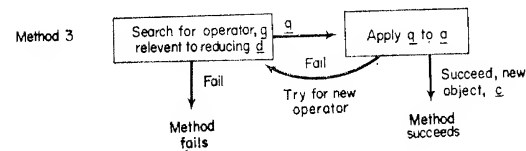


FIG. 5. Means-end analysis.

Figure 5 shows a basic system of methods for what we usually call means-end analysis. It depicts only a core system of inter-linking methods used in GPS. Other methods are known but will

not be discussed here. The method associated with a *transform* goal (Type No. 1) consists in matching the two objects, a and b ; discovering a difference, d , between them (if there is no difference, the problem has been solved); establishing the Type No. 3 goal of reducing d in a ; if this is accomplished, producing c from a , establishing the new transform goal of changing c into b . If this last goal is achieved, the original Type No. 1 goal is achieved.

The method associated with a *reduce* goal (Type No. 3) consists in searching for an operator q , that is *relevant* to the difference d ; if one is found, setting up the Type No. 2 goal of applying the operator.

The method associated with an *apply* goal (Type No. 2) consists in determining if the operator can be applied by setting up a Type No. 1 goal for transforming a into an object $c(q)$, that satisfies the conditions for an input to the operator q . (Generally, an operator can be applied only to objects having certain characteristics. For example, R6 in Fig. 2, can be applied only to a logic expression that has a horseshoe ($\supset$) as its main connector.) If this is successful, the operator q is applied to $c(q)$, producing a new object $p(q)$.

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
Add variables									X	X		X
Delete variables								X			X	X
Increase number			X				X		X	X		X
Decrease number			X				X	X			X	X
Change connective					X	X	X					
Change sign		X			X	X						
Change grouping				X			X					
Change position	X	X										

FIG. 6. Logic table of connections.

The recursive structure of the program is apparent. Transform goals generate Reduce goals and new Transform goals; Reduce goals generate Apply goals; Apply goals generate Transform goals and Apply goals. The additional methods that are known for GPS fall within this same general structure.

We have not yet defined the *differences* that are part of the definition of the Reduce goals, nor the notion of relevant operators. Specific differences and the tests that detect them are not part of the GPS proper but are parts of each particular task environment to which the program is applied. The stub of the table in Fig. 6 is a list of differences that may be detected among logic expressions; a different list might be used for trigonometry, and so on. The columns in Fig. 6 correspond to the operators, also specific to the task, in the logic task environment. The *x*'s in the columns indicate which operators, or rules, are *relevant* to which differences. (How these are determined will be explained later.)

For example, in Fig. 2, the goal of transforming L1 into L4 may lead to detection of a difference in *position* between the two expressions. The operator relevant to this difference (Fig. 6) is R1. Thus, the goal of reducing this difference will generate the goal of applying R1 to L1. Since L1 matches the input to R1, the goal will be achieved, producing L2 and generating the new goal of transforming L2 into L4. Using Fig. 6 and Fig. 2, the reader can simulate GPS's program in carrying through the rest of the solution of this particular simple problem.

The methods just described are ways of setting up sub-goals. In the evaluation part of each goal are tests that allow sub-goals to be rejected as unprofitable, or to be delayed until after more profitable goals are tried.

This is all we shall have to say about the performance program itself. We refer the reader to the other publications already cited, in which we show that GPS will, in fact, solve problems in logic, trigonometry, and elementary algebra, and that certain variants of GPS will simulate in considerable detail the behavior of humans performing the same tasks.

The Learning Problem

When GPS is solving problems in a particular task environment, the performance program consists of two parts: (1) GPS proper, the goal types and methods, which are completely independent of subject matter and are not modified in any way when GPS is applied to a new task environment; and (2) the specifications of the particular task environment: its objects, its operators, and its differences. Within a specified environment, of course, there are many different

problems—proving all logic theorems or all trigonometric identities, for example.

Learning programs might be devised for either main part of GPS. In the present paper we shall only consider programs to enable GPS to improve its performance in a given task environment—to learn heuristics appropriate to that environment. The learning programs themselves will be general—they are programs for learning about any new environment to which GPS might be applied; the content of what is learned, however, will be specific to a particular task environment. These learning programs do not change the core of GPS; instead, they modify the specification of the environment, thus making GPS more efficient in solving problems in that environment.

Learning to characterize in an effective way the task environment is an important and prevalent kind of human learning. A problem solver who is experienced in a particular environment will notice features that will be unnoticed by (or even invisible to) an inexperienced person. The native tracker in the forest is a classical example.

The differences listed in Fig. 6 are features of the logic task environment that are effective for problem solving in that environment. A problem solver who does not have available a good set of differences has little means for working toward his goal except to try, at random or by rote, different sequences of operators until he gets the answer. He cannot even measure progress easily, for the most direct clue to progress is the elimination of differences between the terminal expression and the expressions he has obtained.

After the problem solver has learned to recognize and attend to a useful set of differences, other things remain to be learned about the environment. If a particular difference appears, what operator shall he apply to remove it? He might search the list of operators, again randomly or systematically, until he found one that affected the difference to which he attended. A more efficient procedure would be to build up, once and for all, the table of connections depicted in Fig. 6, which indicates which operators are relevant to the removal of which differences. Equipped with this table, he could, when faced with a difference, consider only those operators from the entire list that are relevant to removing this difference.

It would be possible, in instructing a learner about a new subject, to teach him specifically what differences to attend to and what operators are relevant to what differences: to give him explicitly

a list of differences and a table of connections. In teaching humans we seldom do this. We characterize the task environment by more or less adequate descriptions of the objects and operators (the rules of the game), and perhaps guide somewhat his experiences with the environment. We usually leave it up to the learner to acquire the differences and connections inductively. We assume that humans are equipped with learning programs for improving their performance programs in these two respects. The learner is supposed to be able, himself, to develop a theory about the significant characteristics and structures of the task environment, and to incorporate that theory in his problem-solving program.

In the remainder of this paper we shall discuss in some detail learning programs for the two aspects we have just been considering. The first learning problem we shall pose is: given the operators and the differences in a task environment, to find a good table of connections associating relevant operators with the several differences. The second problem we shall pose is: given the objects and operators in a task environment, and a set of basic tests for discriminating features of objects, to find a good set of differences for that environment. The answers to both questions will take the form of proposed learning programs.

These two aspects hardly exhaust the possibilities for GPS to learn about the particular task environment that confronts it. For example, GPS might learn new operators other than the set given it initially. It might also learn special methods that apply to subclasses of problems. Or it might learn cheap tests to indicate when operators are feasible, thus short-circuiting some of the elaborate general machinery.

Learning the Table of Connections

The learning program required to build a table of connections is quite simple. We describe it because it illustrates a fundamental point about such programs. A simple bit of arithmetic performed on the matrix of Fig. 6 shows that the number of possible tables of connections is not small. There are eight differences and twelve rules in the logic environment, and since each rule might be relevant to each difference, there are $12 \times 8 = 96$ possible connections.

One might use a simple trial-and-error learning scheme to build up the table. By simple trial and error we mean a scheme with the

following general characteristics. (1) Begin with an arbitrary table (perhaps the one that includes all connections). (2) Keep statistics on how often each rule serves to reduce each difference. (3) Try the several rules with frequencies proportional to their relative successes. This procedure incorporates the simplest kind of mechanism of natural selection, to use evolutionary language, or reinforcement of correct responses, to use psychological language.

Most learning programs that have been proposed for computers have this simple character. In some general sense, such learning programs will presumably "work." What is not clear is whether they will work within reasonable time limits in environments of the size and complexity of those we encounter in problem solving.

Such a mechanism might work in the case before us, since the set of possibilities seems fairly regular and small. However, it is not evident that it would be efficient. More important, there is no reason why we should limit ourselves to such mechanisms, which operate entirely inductively from performance, when other information is available. In this case there is information about the structure of the operators which can be used to construct the table of connections directly, without a tedious inductive search.

In the logic environment, each operator is given as a form. Rule 1, for example, will accept as input any expression of the form $(A.B)$ and produce as output an expression of the form $(B.A)$. Now by applying, in turn, the tests for each of the differences to the pair of objects consisting of the input and output form of the rule—that is, to $(A.B)$ and $(B.A)$ —it will be apparent that the only difference between output and input is in the position of the terms. It becomes equally apparent that if Rule 1 operates on another expression, after the latter has been matched to the input form of the rule, it will change only the position of terms in the expression. Hence, the only difference to which Rule 1 is relevant is a difference in position, and the only entry we make in the first column of the table of connections is in the last row.

The remainder of the table can be constructed in the same way. The learning program consists simply in this: consider each operator in turn. Apply, successively, each test for a difference to the operator. If the result of the test is positive, record the operator in question on the list of operators relevant to the given difference. The list of

such lists, over the whole set of differences, is precisely the table of connections.

Let us summarize the analysis briefly. The problem of this particular learning program is to select an element from a set (the set of all possible tables of connections) that satisfies certain conditions. How the learning program can solve this problem depends on what information is available to it. If the program can discover only how the performance program behaves when a given element from the set is incorporated in it, then the learning program can do little more than search blindly and select the elements that work well. If other information is available, however, as in our example, the learning program can incorporate other processes which may be far more efficient than simple trial and error mechanisms. The important empirical question is this: When we consider the learning problems that arise naturally in complex intelligent systems, what is the nature of the information that is available? Is it so scanty as to restrict learning to simple natural selection, or does it allow other, more sophisticated schemes?

With this basic question in mind, we can now examine the second, more complex, learning situation.

Learning a Set of Differences

The second learning situation may be described thus: given the objects and operators in a task environment, and a set of basic processes for detecting and discriminating features of objects, to find a good set of differences between pairs of objects for GPS in that environment.

When we try to define a set of possible differences we discover why this learning task is both difficult and interesting. By a "difference" between two objects, we mean, of course, some characteristic by which they can be distinguished. For the performance program to make use of differences, these must be incorporated in the program in the form of tests that make the appropriate discriminations. For example, for the program to detect that two logic expressions have different connectives, there must be a test that compares the two connectives and records them as identical or different. These tests are, of course, sub-routines in the performance program, and the learning program must be capable of constructing such sub-routines—of writing at least a specialized class of programs.

Unlike the earlier situation with the table of connections, we are here not given the set of possible aspects in any natural way. Before we can discuss the learning program for differences, we must define a programming language from which difference programs can be generated. The programming language must be rich enough to allow adequate learning potentialities; yet simple enough so that the learning program can construct viable routines. And it must not be simply a list of differences already provided for the learner to try. We turn now to the construction of such a programming language.

Programming language for differences. A program to test for a difference must be built up out of some set of more elementary processes that are assumed already available to the learning program for assembly. Prominent among these processes will be a set of primitive discriminations. In some sense they must be more elementary than the differences eventually needed, or the suspicion will remain that the important learning occurred in the selection of these primitive notions and not in the assembly of a program from them. Two devices are available to avoid circularity. We can use a very small set of primitive notions; and we can insist that these same notions be applicable to more than a single task environment. For example, we might start with a computer machine code, and require that all differences be programmed in it for all environments.

We have tried to combine both of these criteria in a Difference Program Language (DPL, for the purposes of this paper). DPL has two parts: there is a general part that consists of a small number of processes for operating on sets and lists. This part is to be used for all task environments, and contains no information about the nature of particular task environments. The second part consists of a list of processes particular to each task environment. These processes constitute basic manipulations and perceptions about the objects in the task environment and form the totality of information about the environment. This list must be given *de novo* for each environment, since we do not assume that there is a common field (as, in perception, the visual field) in which all objects from the several environments are presented. That is, we proceed as if GPS has a different "sense modality" for each environment, and hence must abstract from this into representations of objects belonging to

TABLE 1. PROCESSES FOR SYMBOLIC LOGIC ENVIRONMENT

Symbol	Name	Input	Output	Examples
t	Find terms	object	set of all subobjects	$t(PvQ) = \{PvQ, P, Q\}$
l	Find left	object	left-hand subobject (ϕ if doesn't exist)	$l(PvQ) = P, l(Q) = \phi$
r	Find right	object	right-hand subobject (ϕ if doesn't exist)	$r(Pv - Q) = -Q$
c	Find connective	object	main connective of object (ϕ if object is variable)	
s	Find sign	object	main sign of object	$c((P \supset Q)vR) = v$
v	Find variable	object	variable letter (ϕ if a compound object)	$s(-R.P) = -$
b	Test if constant	anything	input, if constant (ϕ if it contains free variables)	$s(-RvP) = +$
f	Test if free variable	anything	input, if a free variable (ϕ if not)	$v(P \supset Q) = \phi, v(-Q) = Q$
$\supset$	Test if $\supset$	connective	$\supset$ if connective is $\supset$ (ϕ if not)	$b(PvQ) = PvQ, b(A) = \phi$
v	Test if v	connective	v if connective is v (ϕ if not)	$f(A) = A, f(P) = \phi$
$.$	Test if $.$	connective	$.$ if connective is $.$ (ϕ if not)	$\supset(\supset) = \supset, \supset(.) = \phi$
$+$	Test if $+$	sign	$+$ if sign is $+$ (ϕ if not)	$v(v) = v, v(\supset) = \phi$
$-$	Test if $-$	sign	$-$ if sign is $-$ (ϕ if not)	$.(.) = ., .(v) = \phi$
$A(B, C, \dots)$	Test if $A(B, C, \dots)$	variable	A if letter is A (ϕ if not)	$+(+) = +, +(-) = \phi$
$P(Q, R, \dots)$	Test if $P(Q, R, \dots)$	variable	P if letter is P (ϕ if not)	$-(-) = -, -(+)= \phi$
				$A(A) = A, A(P) = \phi$
				$P(P) = P, P(Q) = \phi$

TABLE 2. GENERAL DPL PROCESSES

Symbol	Name	Input	Output	Examples
$A[X]$	Assign	any object	X if input is not ϕ , ϕ if input is ϕ	$A[+] (P, Q) = +, A[+] \phi = \phi$
$\bar{A}[X]$	Inverse assign	any object	X if input is ϕ , ϕ if input not ϕ	$\bar{A}[+] (P, Q) = \phi, \bar{A}[+] \phi = +$
$B[X]$	Blank	any object	goes through all subparts of input: x .	
C	Find component	set	If $X(x)$ not ϕ , replace $X(x)$ by ϕ in input	$B[c] (P \supset (Q \vee R)) = P \phi (Q \phi R)$
D	Difference	any pair of objects	takes any component for output	$C[\supset, \vee, \wedge] = \vee$
E	Expand	set of sets	compares corresponding subparts of the two inputs. If equal, replaces each by ϕ . Output is modified pair.	$D[(P, Q, R), (Q, Q, R, P)] = ((P, \phi, \phi), (Q, \phi, \phi, P))$
F	Find first	list	Output is set of all elements in the subsets of input, with multiplicity	$E\{(P, Q)\}; \{P, R\} = \{P, Q, P, R\}$
$G[X]$	Group	set	first item on the list (ϕ if doesn't exist) output is a set of sets. Each subset contains all the items of the input set with the same value of $X(x)$	$F(P, Q, R) = P$ $G[I] \{P, P, Q, P, R, Q\} = \{\{P, P, P\}, \{Q, Q\}, \{R\}\}$
I	Identity	any object	input	$IX = X$
$K[X]$	Constant	any object	X , for any input	$K[+] P = +$
L	Find last	list	last item on the list	$L(P, Q, R) = R$
M	Find prior	item from list	item preceding input item (ϕ if doesn't exist)	MP from $(Q, R, P, S) = R$
N	Find next	item from list	item following input item (ϕ if doesn't exist)	NP from $(Q, R, P, S) = S$
P	Intersection	set of sets	set of items common to all subsets of input	$P\{(P, Q), \{Q, R\}, \{P, Q, R\}\} = \{Q\}$
$R[X]$	Select representative	set	set consisting of one representative element of each value of $X(x)$. Compare $G[X]$	$R[I] \{P, P, Q, P, R, Q\} = \{P, Q, R\}$
$S[X]$	Select	set	set of items of input set with $X(x)$ not ϕ	$S[\vee] \{P, Q, -R, R \supset P, R\} = \{-R, R\}$
U	Set	list	set of items on list	$U(P, Q, R) = \{P, Q, R\}$

M

the general part of DPL before it can describe these objects in terms of common properties or conventions.

Table 1 gives a list of processes for the environment of symbolic logic. The symbol, ϕ , which represents the null set, is used to record that a test was negative or that a find process had a null output. Note that none of the processes of Table 1 can be omitted (without some equivalent replacement) if the set is to be complete for logic expressions. If one or more processes were deleted from the list, and the remaining processes were our only source of information about features of logic expressions, we could never become aware of the missing elements. Each process accepts only certain types of inputs and produces specified types of outputs, as shown in the figure. Besides the list of processes, a list of input and output types is provided. Thus c , the process that finds connectives, can be applied only to objects (logic expressions), and produces a connective, which is a symbol of a different type.*

The general part of DPL consists of the seventeen processes shown in Table 2. The inputs and outputs of these processes are lists and sets (unordered lists) of items. With two exceptions, the processes treat objects of the task environment as unanalyzable units. Thus, the general part of the language is independent of information about particular task environments. One of the exceptions is the process $B[X]$, whose operand may be an object, set or list. This process replaces each component of a specified kind in the input by the null symbol, ϕ . For example, $B[c]$ replaces all connectives in an object by ϕ , so that the latter symbol now serves as a generalized abstract connective. Thus $B[X]$ must be able to "get at" all the subobjects of the object on which it operates.

The other process that requires information about the structure of task environment objects is D , a process that finds differences between pairs of objects. Since this process is of central importance, it deserves extended discussion. The input to D is a pair—that is, a list of two items, say X and Y . The members of the pair (X, Y) may be objects or they may be sets, lists, or list structures. The output of D is also a pair of items, say (X', Y') , obtained as follows.

* This information is implicit in the operation of the processes, and could be learned by GPS by a program not very different from that for learning the table of *connections*.

(1) X and Y are put into correspondence according to their structure. For example, if X and Y are logic expressions, they are lined up with their main expressions together, left-hand subexpressions together, right-hand subexpressions together, and so on.

(2) Any corresponding parts of X and Y , respectively, that are identical are replaced, in both X and Y , by ϕ . When this process has been completed for all pairs of parts of X and Y , a new pair of objects (X', Y') , will have been obtained in place of (X, Y) . The new pair (X', Y') , is the output of the process D . In this new pair, X' consists of all parts of X not belonging to Y , and Y' of all parts of Y not belonging to X .

The processes are the elementary terms of DPL; we must also provide ways for compounding programs of them. Speaking roughly, DPL programs are sequences of DPL processes. Sequences of processes, each term operating on the output of the preceding ones, are written horizontally. The operation proceeds from right to left as in standard mathematical operator notation. Apart from simple sequences, four other combining operations are needed. These are indicated in Table 3. For example, the notation permits us to write down a set of processes as though it were a simple process. The output of this set is the set of outputs that would be produced by each of the component processes, operating independently on the input. These various modes of combination of

TABLE 3. RULES OF COMBINATION

Let P and Q be processes and X and Y be operands.
$P.X$ or $PX =$ apply P to X
$P*\{X, Y\} = \{PX, PY\}$
$P*(X, Y) = (PX, PY)$
$(P, Q):(X, Y) = (PX, QY)$
$\{P, Q\}X = \{PX, QX\}$
$(P, Q)X = (PX, QX)$

processes are the functional equivalent, in DPL, of such features normally found in programming languages, as iterations, conditional transfers, and working storages.

DPL consists, then, of a set of basic processes and some ways of combining processes. It is not a complete programming language; it omits several important notions, such as ordering relations and

recursive definition, in the interest of simplicity. However, DPL is adequate for constructing a rather large set of differences, including those we have used in the performance program of GPS for symbolic logic.

Figure 7 shows, by a step-by-step analysis of an example, how the DPL program will find the difference between the sets of variables contained in two logic expressions. The input (X, Y) , to the program is the pair of logic expressions $(P \cdot Q, Q \supset Q)$. The parenthesis followed by the asterisk (*) indicates that the whole sequence of operations, $R[I]vt$, is to be applied to each member of the pair, that is, to X and to Y separately. Then, the differencing operation, D , is to be applied to the resulting pair of outputs. First,

$D(R[I]vt)^*$	(X, Y)
	$(Q \supset Q, P \cdot Q)$
t^*	$\{Q, Q \supset Q, Q\}, \{P, P \cdot Q, Q\}$
v^*	$\{Q, Q\}, \{P, Q\}$
$R[I]^*$	$\{Q\}, \{P, Q\}$
D	$\{\phi\}, \{P\}$

FIG. 7. Difference in set of variables.

application of t to each logic expression produces, for each, a set whose elements are the subobjects included in the original expression (including the expression itself). In our example, each set produced by t contains three elements. Next, application of v to each of these sets replaces each of its elements that is compound—that is not a variable—by ϕ , and retains the variables. Next, application of $R[I]$ eliminates all multiple occurrences of identical symbols—e.g. it reduces (Q, Q) to (Q) . Finally, application of the difference, D , reduces the left-hand set to P and the right-hand set to ϕ . This means, as it should, that the left expression in the original pair

contains the variable P , which does not occur in the right expression, but that all variables in the right expression occur in the left.

The most striking characteristic of differences written in DPL is that they are "abstractive." They start with comparisons of the full detail, and gradually remove distinctions by successive application of DPL operations.* This is to be contrasted with normal programming tests, which are "discriminative," in that each elementary process only discriminates a very minute part of the object, and more and more information is built up about the objects by constructing a tree of tests.

Learning situation for differences. Having described a relatively rich programming language having a simple structure, which expresses easily some kinds of differences we know to be useful, we can now return to the problem of how a set of differences appropriate to a particular task environment might be learned. The problem can be reformulated thus. Given the objects, operators, and list of basic designations for a task environment, to find a good set of differences, expressed in DPL, to be incorporated in GPS for effective problem solving in that environment.

It would not be hard to design such a learning program based on the simple trial-and-error prototype. The program would generate sequences of processes in DPL, and test these for their usefulness as differences. Because DPL has a simple structure, most such sequences would be viable programs, and perhaps a significant proportion of them might even be interpretable, in some sense, as differences. The learning program would incorporate such sequences, tentatively, in the performance program and keep statistics on their application in successful attempts on problems. If a difference were tried and found wanting, the learning program would remove it to make way for a new candidate. Gradually, in the fullness of time and with a nonhostile task environment, the learning program might evolve a satisfactory set of differences.

This much can be done, but we have no reason for *a priori* confidence—or even hope—that the learning will be accomplished

* DPL is similar in this respect to the language constructed by Selfridge and Dinneen^(6,7) for a pattern recognition program. Using the alternative approach Mr. Edward Feigenbaum of Carnegie Institute of Technology, in a forthcoming report, describes a system of discriminative tests that comprise part of a program for simulating human rote learning.

within a reasonable time span. The space of possible differences is very large, and humans guide their trial-and-error searches through it with a variety of heuristics. If the learning program is to operate in real time, it must make use of additional information in selecting and testing candidates for the set of differences. There is, in fact, a great deal of information available in the task situation, if the learning program can get access to it: information about the structure of the operators; information about the elementary processes of DPL; information about criteria for a good set of differences; information about the transformations that particular operators produce on particular objects in the task environment; and so on. The learning program might even conduct investigations to obtain additional information: for example, it might explore the designation processes in the task environment to discover their mutual relations.

If it is to use such information fruitfully, the learning program must be intelligent—it must be a problem-solver. It is doubtful that a simple process, like the one we used to construct the table of connections, exists in this situation. A learning program that resembled a chess-playing or theory-proving program would have a better chance of succeeding. At any rate, this is our hunch: that an intelligent learning program will differ from a problem-solver, not in its structure, but only in the content of its task.

Since GPS is a problem-solving program having pretensions of generality, we can try to use GPS itself as the learning program. If we can restate the learning problem as a problem involving the application of operators to objects in order to remove differences, then, upon presentation of a suitable goal, GPS should be able to work on the new (learning) problem of creating good sets of differences for the original task environment, and should be able to bring to bear on this learning problem its full repertoire of heuristics.

The idea of using a single “intelligent” program to bootstrap itself appeals to deeply-rooted notions about the reflexiveness that is involved in self-organization. But the strategy has other attractive features. First, it provides a test of the power of GPS and a source of ideas for expanding its repertoire of goals and methods. Second, if the program for learning on a single aspect of performance is as large and complex as the performance program itself, only by using the same program in both roles can we hope

to keep the size of the total system within tolerable bounds. This argument becomes even more compelling when we consider the problems of learning on all the other aspects of each performance program.

Our task, then, in the remainder of the paper is to attempt to translate this learning situation into GPS terms, and to evaluate the chances that GPS can handle the learning problem successfully. We will proceed by setting up each of several GPS task environments that seem to be required, and defining the objects, operators, and differences in them. Our own goal, in exploring this path, was to create enough mechanisms to allow us to hand simulate GPS in the process of learning differences. Thus we could provide some assurance that all the essential parts had been identified. We achieved this, but at the cost of a large amount of detail. In the pages that follow we will give just enough of this detail to convey the general form of the solution and to allow a meaningful sketch of the hand simulation. Since this is not nearly enough information to allow anyone else to verify what we have done, we put the scheme forward in a very tentative spirit.

Basic Task Environment for Learning

From now on we will be applying GPS to several task environments. Since all of them will be formally similar—involving objects, differences, and operators—we need to label them if we are to avoid confusion. We will introduce two of these, the *A*-environment and the *B*-environment, at the outset.

The A-environment. We shall call the original task environment (e.g. logic) the *A-environment*. The *A*-environment will have *A*-objects (logic expressions), *A*-operators (rules), and a list of *A*-designations (e.g. the test for a connective). The learning problem is to find a good set of *A*-differences (like the set in the table of connections).

The B-environment. We shall call the environment of the initial learning problem (to learn a good set of *A*-differences) the *B-environment*. The *B*-environment will have *B*-objects (sets of *A*-differences), and the learning problem is to find a *B*-object that makes for good problem solving in the *A*-environment. Our task is to create *B*-operators (operators for creating and modifying sets of *A*-differences), and *B*-differences (tests for comparing *B*-objects),

and to discover how to state the learning goal as a *B*-goal. The *B*-differences must be independent of the particular task environment, the *A*-environment, and must use only information derivable from the list of *A*-designation processes, *A*-operators, and samples of *A*-objects.

TABLE 4. THE *B*-ENVIRONMENT*B-Operators*

- Q1 Add an *A*-difference that gives + for pair *X* and ϕ for pair *Y*. (A pair may either be a pair of objects, or the condition and product forms of an operator.)
 Q2 Modify *A*-difference *T* to give + for pair *X*.
 Q3 Modify *A*-difference *T* to give ϕ for pair *X*.
 Q4 Delete *A*-difference *T* from set *S*.
 Q5 Add an *A*-difference that gives + for pair *X*.

B-Differences

- D1 The set of *A*-differences not consistently defined for some pair of objects.
 D2 The set of *A*-operators with no associated difference.
 D3 The set of *A*-object pairs with no associated difference.
 D4 The set of non-orthogonal situations (each situation consists of an *A*-object, a list of *A*-operators, the product from applying the operators to the given *A*-object, and the new differences between the input and output that are not associated with any of the operators).
 D5 The set of full *A*-differences (having all *A*-operators associated with them).
 D6 The set of empty *A*-differences (having no *A*-operators associated with them).
 D7 The set of *A*-differences with more than one associated *A*-operator.
 D8 The set of *A*-operators with more than one associated *A*-difference.
 D9 The total number of *A*-differences.

Table of connections for B-Operators and B-Differences

	Q1	Q2	Q3	Q4	Q5
D1		x	x	x	
D2	x	x			x
D3		x			x
D4		x	x		
D5			x	x	
D6		x		x	
D7			x	x	
D8			x	x	
D9				x	

The *B*-operators work on sets of *A*-differences. They add *A*-differences to a set, delete *A*-differences from a set, or modify existing members of a set. Since what is available to the learning program are samples of *A*-objects and *A*-operators, *B*-operators are needed that construct *A*-differences on the basis of their behavior for given samples—i.e. that produce *A*-differences defined extensionally. Table 4 gives five such *B*-operators, enough for the purposes

of this paper. Except for Q4, each seeks to obtain an A -difference that gives a specified result, $+$ or ϕ , for a specified pair of objects. (The $+$ is a conventional symbol that means that the A -difference "holds" for the pair of A -objects—that is, gives some non-null output.) These B -operators are applicable to any A -difference in the set, just as in logic or algebra, the commutative law may be applicable to several parts of an expression. The particular element to which a B -operator is to be applied is determined by B -differences, which we shall consider in a moment, or by additional selective heuristics that we shall discuss in more detail a little later. Defining operators by giving the properties of the things they produce does not guarantee that such operators exist, or that, if found, they will accomplish what is wanted of them.

Before we consider the problem of constructing the B -operators, let us look at the B -goals and the B -differences. The highest B -goal for the learning process is a criterion for a good set of A -differences. Ultimately, a good set of differences is one that is effective for problem solving in the A -environment. But to permit intelligent learning, other ways must be found to characterize good sets of differences, so that GPS, in its learning efforts, can evaluate the improvement it is achieving.* We shall provide in the B -environment, the following criteria of a good set of differences:

- (1) Only one or a few A -operators should be relevant to each A -difference in the set.
- (2) Only one or a few differences should be associated with each operator.
- (3) Each operator should be relevant to at least one difference.
- (4) Each pair of non-identical A -objects should exhibit at least one A -difference in the set.
- (5) The set of A -differences should be nearly orthogonal—that is, if only difference D is relevant to a particular operator, then application of this operator to an object should produce only difference D between the input object and the output object.
- (6) An A -difference should always give the same result when applied to the same A -objects.

* We will not consider whether GPS could itself construct the intermediate criteria given only the ultimate performance criteria and experience in several task environments.

The rationale for these criteria is rather simple. The differences are the diagnostic tests that GPS uses to determine what operators it should apply in a given situation. The diagnosis will be most efficient if each difference points to the application of one and only one operator, if each operator affects one and only one difference, if the effect of an operator is predictable, and if a difference is always detectable between nonidentical objects. In the limiting case, with a "perfect" set of differences, the performance program would have the trivial task of finding the differences between a given and a desired object, and applying, in sequence the (unique) operators for removing the several differences. Of course, in general, no such perfect set of differences exists or can be found; the task of the learning program is to approximate it as closely as possible. Moreover, there may be more than one satisfactory set of differences. Any such set is a theory of the important features of the task environment and their interrelations.

B-differences, in the light of this discussion of goals, are simply features of sets of *A*-differences that describe in what respect those sets meet or fail to meet the above criteria. Table 4 states these differences in measurable form, and gives the table of connections between the *B*-differences and the *B*-operators. To weld all the separate "reduce difference" goals in the *B*-environment into a single effective goal, we establish a priority order of the differences, giving highest priority to completeness properties. (They are so ordered in Table 4.) GPS will be instructed to produce a set of *A*-differences, attending first to consistency and completeness requirements. Once these are satisfied, GPS will attempt to improve the set with respect to the lower-priority criteria, always returning to the higher criteria if these are no longer satisfied after the set has been modified.

Let us summarize what we have said up to this point about the learning task. The task is defined in the *B*-environment, which is an environment suitable for GPS. The *B*-objects are sets of differences in the *A*-environment; the *B*-operators, shown in Table 4, permit manipulation of these sets; and the *B*-differences, also shown in Table 4, correspond to various criteria for "good" sets. The learner's goal is not to attain a fixed, given *B*-object, but to construct a series of *B*-objects in an attempt to reduce the *B*-differences. We have left open the problem of how the *B*-operators

are to provide differences satisfying the criteria specified. The *B*-operators assume there is some effective way of programming in DPL to provide suitable *A*-differences. We now turn to this problem.

Task Environment for A-Differences

If GPS is to construct DPL programs for *A*-differences, then programming in DPL must be described as a GPS-type task. Again, we need an environment in which this task can be performed.

The C-environment. The natural environment for the task is one where *C*-objects are DPL programs. Then the *C*-operators are ways of putting programs together; the *C*-differences are things that can be noticed about programs, such as whether one contains a *D* (difference) or not, and the *C*-goals are set up by the *B*-operators: to transform the basic set of programs (the given *C*-objects) into a program (a new *C*-object) with certain features.

However, a different environment may be considered. Programs consist of sequences of subprograms—ultimately of sequences of the primitive DPL processes. A program takes an input and transforms it step by step until it finally is made into the output. This sequential character suggests an environment where the objects are the various inputs and outputs, and the operators are elementary DPL processes. Then the final desired program is the sequence that transforms an initial input to a final output.

The situation here is exactly analogous to that in theorem-proving. In symbolic logic, for example, the problem is stated: Prove theorem *T*, given axioms *A*, *B*, What is wanted is a proof. But instead of working in the space of proofs—that is, in an environment where proofs are objects—one works in an environment where logic expressions are objects, rules of inference are operators, and the desired proof is the sequence of rules that is applied to get from the axioms (the given objects) to the desired theorem (the final object).*

* Seen in this light, the original LT program⁽⁵⁾ was itself a program writer, which generated a program—the sequence of methods leading to the proof—that would produce logic expressions from other logic expressions. It was, of course, an uninteresting programmer, since the product has no particular usefulness as a program, but it provides a direct model for the present self-programming scheme.

We will work entirely in this latter environment, which we will call the *D*-environment. Our reason for mentioning the *C*-environment is to explain the ways in which GPS must be extended to work in this *D*-environment. Not all the relevant information can be obtained by examining the inputs and outputs of programs. An important part of the problem-solving information comes from the properties of programs, viewed as objects (e.g. whether the program contains some task-environment processes or only general processes). Thus GPS must consider not only differences between inputs and outputs, but features of the sequence it is building to bridge the gap—that is, differences that properly belong to the *C*-environment. The situation is again analogous to theorem-proving where one may impose such constraints as finding the shortest proof, finding an elegant proof, or finding a proof using a given theorem.

The D-environment. Let us define the *D*-environment more carefully. *D*-objects are the inputs and outputs of DPL programs. *A*-objects are included, since these form the initial inputs to the *A*-differences. All the intermediate products are also included among the *D*-objects: *A*-objects with parts replaced by ϕ 's (from $B[x]$), sets and lists of *D*-objects, the symbols ϕ and $+$, and so on. No circularity arises from inclusion of the *A*-objects, since the only information available in the *D*-environment about the *A*-objects is that already available to GPS for learning: the list of environmental processes with their input and output types as shown in Table 1.

By definition, the *D*-operators are DPL programs: they transform *D*-objects into other *D*-objects. The set of *D*-operators includes all of the DPL primitive processes, as given in Tables 1, 2 and 3—some 42 operators in all. It will be noticed that a number of operators contain free variables, whose values are other operators. This is true, for example, of the five operators of Table 3, which only specify ways of combining other operators.

Beside the *D*-operators indicated above, we need operators that allow other manipulations of DPL programs than merely adding a new operator to the front end of a sequence of DPL operators. These additional operators are *C*-operators, properly speaking. However, for the simulation only two such operators were required: one that deleted the last *D*-operator of a sequence, thus going back

"one step"; and one that deleted a D -operator satisfying certain conditions from the middle of the sequence.*

The D -differences are based partly on features of D -objects. GPS can only detect features of D -objects within the limits of the information available for learning. Considered as a D -object, an A -object is an unanalyzable unit, and the D -differences must treat it as such. GPS is able, of course, to examine the sets and lists of objects and parts of objects that are created in the course of a DPL program. Nevertheless, the features of D -objects it detects are rather general.

Type. GPS can tell whether it is working with an object, a set of objects, a set of sets of connectives, and so on. A glance at the D -operators in Tables 1 and 2 shows that D -operators vary considerably as to the type of D -object taken as input, and the type produced as output.

Size. It is possible to count how big a D -object is, taking as the unit the innermost component. Thus GPS can assign size 3 to a list of three objects, or size 21 to a set of seven sets of three variables each, and so on. It cannot assign a size measure to single A -objects, of course, and must treat them all as equivalent.

Variety. It is also possible to measure the variety of a D -object—to count how many *different* things there are in the D -object. This is possible because the process, D , provides a test of identity. Variety is a useful notion because certain D -operators, such as $B[X]$, decrease the variety without changing the size of a D -object.

Sign. Finally, GPS can determine whether a D -object is $+$ or ϕ . (Formally this can be done by applying the operator $A[+]U$ to the D -object. This produces the output $+$ unless the object is ϕ , $(\phi, \phi) \dots$, in which case it produces ϕ .)

The D -differences are also based partly on features of the DPL programs that produce a D -object. Again, these features lead to C -differences, properly speaking. The features we will need are the following.

* Constructing programs by "working forward" in the D -environment, adding one process at a time, is directly analogous to the way by which the completed DPL program will carry out its information processing. This "analoging" is a very common human technique for programming. But this is not the only consideration that goes into human programming, and we suspect that for programming tasks more complicated than our elementary example, other C -operators will be needed.

Set of environmental processes. GPS can note which of the processes that occur in the DPL program also occur on the list of environmental processes given GPS to characterize a particular A -environment (Table 1).

Set of general processes. GPS can note, similarly, what general processes occur in a DPL program.

Set of special processes. GPS can tell if the program contains some special process, like the constant operator $K[X]$.

Contains a D . GPS can note whether the program contains a D anywhere. This is a very important feature, since this one operator takes a difference between two objects, and thus must be a constituent of every A -difference.

Consistency. A number of DPL processes involve selections from sets—e.g. C , which simply selects a member of the set to which it is applied. Which member is selected, within the constraints laid down by the process, is a matter of “chance.” It often happens that the final output of a program is critically dependent on such a chancy event. Thus an A -difference may sometimes give $+$, sometimes ϕ when applied to the same pair of A -objects. We assume GPS can detect such inconsistencies.

Given these features of D -objects and DPL programs, we can construct D -differences, based on the ability of the various D -operators to change one feature into another. Instead of laying out the table of D -differences, which is rather large and complicated, we will content ourselves with indicating, as we discuss the simulation, those D -differences that played an important role.

It remains to describe the topmost D -goals, which are set up by B -operators. Consider the B -operator Q5: Add an A -difference that gives $+$ for the pair A . In the D -environment this can be phrased as: Transform A into $+$. However, there are some important side conditions. First, the DPL program that transforms A to $+$ must be a difference. This can be expressed by requiring the program to contain a D .* Second, the program should not be

* An appropriate operational definition of “difference” should also require that the two inputs to D be dependent on the two input A -objects, and that the output of the program be dependent on the output of D . This dependency can be measured structurally by drawing the oriented graph corresponding to the information flow through the program. Since these conditions never affected the simulation, we indicate them only in passing.

trivial. After all, the constant function will yield $+$ if applied to an A -object. We can give a partial list of excluded special programs. Taking these points into account, a more complete formulation of the D -goal corresponding to Q5 would be: Transform A into $+$, in such a way that the transform is a difference and is not trivial. This goal implies a slight generalization from the form of the transform goal given at the beginning of the paper. There we said: transform object A into object B . Here we allow ourselves to require that additional conditions be satisfied. There is no difficulty in doing this, however, as long as GPS can recognize the existence of unsatisfied conditions and can set up differences and reduce goals based on them.

A glance at the other B -operators shows that similar formulations hold for each one, except Q4, which has no side conditions. Q1 requires a single transform that accomplishes two transformations simultaneously. This is important, since Q1 is the B -operator that brings about a discrimination. Both Q2 and Q3 require that the transform goal start with a partially-completed sequence, although modifications are allowed in this initial segment.

The D -environment is now complete. It differs sufficiently from the initial environments of GPS to require some additional heuristics. The need arises from the large number of D -operators that satisfy various differences. This multiplicity of operators makes further subselection both necessary and profitable. The selection is accomplished in two ways:

- (1) In selecting an operator, GPS will also consider *feasibility*; that is, it will match the input type of the operator to the type of the object.

This added test will reduce the attempts to fit infeasible operators. This is reasonable in a situation in which feasible operators are always available.

- (2) If more than one operator is available after selection by a series of criteria, then the original goal will be consulted and the next most important difference will be generated to provide an additional means of selection.

Thus at each search for a D -operator selection may take place on a number of criteria. Given this opportunity for multiple selection, we can influence the construction of DPL programs by adding further conditions to the goals set up by the B -operators. Besides

requiring that the final program be an *A*-difference, we can also require that it contain some task-environment processes, but that it have few of these in common with the other difference programs already in the set. These requirements are distinctly heuristic, for their aim is to bias the order in which programs are constructed. By giving the heuristic conditions low priority we allow their use occasionally as selective principles when there is lots of choice available, but assure that they will not override more crucial conditions.

Simulation of Learning

We have now described the task environments that will let GPS work on the problem of learning its own sets of differences. This has required a considerable amount of specification: a new programming language, and three GPS environments. And, although we have been fairly specific in describing the language, operators and differences, a number of gaps still exist. Hopefully, however, these are gaps of detail and the essential mechanisms have appeared.

To shed some light on the question of completeness—which is crucial in an initial exploration—we tried to simulate the program by hand. We took logic as the *A*-environment, for which GPS was to produce a set of *A*-differences, corresponding to part of those in Fig. 6. The course of this simulation is given in Table 5. It is very crude. Answers to many questions of detail were created on the spot during the simulation. Many arbitrary selections were made, often without a formal scheme.

The simulation was based on the simple example used earlier to illustrate GPS (Fig. 2). The four operators that were chosen, R1, R2, R5, and R6, are the ones involved in that example; and when a sample problem was needed at step 4 of the hand simulation, that was the one used.

Starting from scratch in Step 1, GPS used operator Q1 to insure that the resulting *A*-difference would discriminate *something*. The exploration in the *D*-environment is shown for this goal, to find an *A*-difference that is $+$ for R1 and ϕ for R2. What occurred was rather simple. The initial program, consisting simply of process *D*, was obtained because of the high priority given (by the *D*-differences) to programs containing process *D*. Each partial sequence, *D*, *LD*,

TABLE 5. SKETCH OF HAND SIMULATION WITH LOGIC AS THE A-ENVIRONMENT

The set of A -operators for which A -differences are to be found is:

R1: $A \vee B \Rightarrow B \vee A$ or $A \cdot B \Rightarrow B \cdot A$

R2: $A \supset B \Rightarrow \neg B \supset \neg A$

R5: $A \vee B \Rightarrow \neg(\neg A \cdot \neg B)$ or $A \cdot B \Rightarrow \neg(\neg A \vee \neg B)$

R6: $A \supset B \Rightarrow \neg A \vee B$ or $A \vee B \Rightarrow \neg A \supset B$

The initial set of A -differences is the null set:

S0: ϕ

1. The set, S0, is operator incomplete (see D2 in Table 4) since none of the operators now have associated differences. The first goal given the D -environment is to find an A -difference that is $+$ for R1 and ϕ for R2. The following exploration is conducted in the D -environment:

$\nearrow DD$ (reject)

$D \rightarrow FD$ (reject)

$\nearrow S[+]s^*R(c)tLD$ (reject)

$\searrow LD \rightarrow tLD \rightarrow R[c]tLD \rightarrow s^*R[c]tLD \rightarrow S[-]s^*R(c)tLD \rightarrow$

$\rightarrow \bar{A}[+]S[-]s^*R[c]tLD$

This gives the next set, S1, of A -differences:

S1: T9 = $\bar{A}[+]S[-]s^*R[c]tLD$

R1	R2	R5	R6
+	ϕ	ϕ	?

2. T9 is inconsistent with R6—it gives $+$ or ϕ depending on arbitrary selective processes. The next goal in the D -environment is to modify T9 to produce $+$ for R6. This is accomplished, giving the next set:

S2: T11 = $S[-]s^*tLD$

R1	R2	R5	R6
ϕ	+	+	+

3. R1 is no longer covered by S2, so the next goal in the D -environment is to create an A -difference that is $+$ on R1 and ϕ on R5. This leads to the set, S3:

S3: T11 = $S[-]s^*tLD$

T15 = $A[+]D(l, r) : D$

R1	R2	R5	R6
ϕ	+	+	+
+	ϕ	ϕ	+

4. With S3, all operators have associated A -differences, and all the A -differences are consistent. In order to see if S3 could distinguish non-identical pairs of objects, the simple problem of Fig. 2, Transform $S.(-P \supset Q)$ into $(Q \vee P).S$, was attacked with the above table of connections. T15 was $+$ for the pair of objects, and R1 was applied, just as in Fig. 2. However, no difference was found between the left side of L2 and the left side of L4, even though they are not identical. The next problem for the D -environment was to find an A -difference that would produce $+$ for this pair, $(-P \supset Q, Q \vee P)$. This resulted in the next set:

S4: T11 = $S[-]s^*tLD$

T15 = $A[+]D(l, r) : D$

T20 = Dc^*

R1	R2	R5	R6
ϕ	+	+	+
+	ϕ	ϕ	+
ϕ	ϕ	+	+

S4 distinguishes all the pairs of non-identical objects generated in solving the test problem. The simulation was terminated at this point.

tLD , and so on, produced $+$ with both R1 and R2. Therefore, new processes were added that "decreased size" or "decreased variety" until finally, with the addition of $S[-]$, a program was produced that gave ϕ with R1 and $+$ with R2. This output was just the opposite of what was needed; but the discrepancy was detected by a "reversal" difference and the routine $\bar{A}[+]$ was applied to change the $+$ to ϕ and the ϕ to $+$. The various branches that were generated but not further explored were rejected either because the output was identical with the input, thus indicating no progress, or because the outputs from both R1 and R2 were identical, so that the process did not discriminate between these two rules.

The question of consistency, which seems a rather technical detail, produced the next phase of the simulation. When T9 was applied to R6 it sometimes gave $+$ and sometimes ϕ . When operator Q2 was invoked to remedy this, it produced two applications of C-operators. The first of these deleted $R[c]$ from the middle of T9, where this process had been identified as the culprit causing the inconsistency. $R[c]$ was identified by tracing through the flow of information. Deletion of $R[c]$ yielded a program (T10) that was ϕ on R6. This discrepancy was detected by the "reversal" difference and the $\bar{A}[+]$ was stripped off the front of T10, giving T11 an A -difference that satisfied the goal.

The change from T9 to T11 left R1 not covered by any A -difference. Q1 was again applied to get an A -difference that would be $+$ on R1. Since a second pair of objects was needed to specify Q1 completely, the input and output forms of R5 were chosen arbitrarily as the second pair. The result of the problem-solving in the D -environment was T15. This test was developed in a similar manner to that which produced T9. However, the heuristic of choosing different task-environment processes for the two tests resulted in the occurrence of r and l as components of T15. Again the "reversal" difference accounts for the $\bar{A}[+]$ at the front of T15.

The new set of differences, S3, had all operators covered. It was necessary, next, to use some A -objects in order to test whether S3 could discriminate pairs of non-identical objects. Using the sample problem of Fig. 2, a pair of objects, $(-P \supset Q)$ and $(Q \vee P)$, was found that were not identical, but still yielded ϕ when T11 and T15 were applied to them. This caused the next problem-solving attempt

in the *D*-environment, defined by operator Q5: to discriminate these two objects. T20 was found to do this. It was added to the set to make S4. This final set proved satisfactory for the remainder of the sample problem, and the simulation was terminated. If simulation had continued, either more problems would have been generated to test for object coverage and orthogonality, or some lesser differences would have been tried in order to improve the discriminability of the table.

CONCLUSION

We conclude this paper by listing some observations—both reassuring and discomfiting—on the path we have followed.

(1) The rough simulation presents good evidence, we think, that we have specified the mechanisms that are essential to permitting GPS to work on its own learning. These mechanisms delineate at least one variety of intelligent learning.

(2) A feature that stands out clearly in the program is the interaction between the two environments, *B* and *D*, one providing the goals for the other. This makes good sense in the light of our general knowledge of computer coding. The distinctions commonly made between “programming” and “coding,” and between “problem-oriented languages” and “machine-oriented languages,” may reflect the relation between the two environments.

(3) An interesting feature of the learning task is that the set of differences is a very ambiguous object. No difference can be completely evaluated in isolation, since the properties of the set as a whole determine how effective the problem solving is. Similarly, complete factorization with each operator associated uniquely with a particular difference seems unattainable. Hence, the goal of obtaining a satisfactory set of differences, unlike GPS goals considered previously, is not a search for a unique specified object. This difficulty was circumvented by the form of the Reduce goals (D1 to D9), which give GPS “direction” in its learning task without specifying a definite final resting place.

(4) A considerable amount of mechanism has been added to GPS, but none of this new mechanism seems peculiar to the learning of logic, and much of it, such as the additional selection mechanism, is of the same generality and spirit as the initial version of GPS.

(5) In spite of this apparent generality, the justification for much of the mechanism rests on the possibility of using it for a number of different task environments. This possibility is quite untested, for all our work here has been limited to the task of logic. Although GPS has pretensions of being general, only two task environments (logic and elementary algebra with trigonometric functions) have been specified with sufficient precision to exhibit in detail the set of differences. Hence a question that occurs prior to testing this learning program is whether a sufficient population of environments can be constructed in which GPS can operate.

(6) A more serious problem is that a general set of differences may possibly exist that would be effective for all environments. Some of the differences in the logic situation—the set of different kinds of things, and identity of symbols of a given class, like connectives—have a very basic ring. In trigonometry, the one other environment for which we have good information, exactly the same set of differences was used as for logic, with the exception of commutivity and associativity, which were incorporated in the structure of the objects as in normal algebraic notation. If such a “universal” set of differences existed, it might still leave the task of applying differences to the specific environment, but this is a task more like learning the table of connections than like the learning task we have just analyzed. Supporting this possibility is the fact that it is difficult to imagine how differences like commutivity could be built up from more elementary notions. In the simulation this is achieved by the *A*-difference T15, which compares the left-hand side of one object with the right-hand side of the other.

(7) A final oddity of the present scheme was noted earlier. Different task environments for GPS are completely independent, much more so than the different task environments for a human, which all occur, in the last analysis, in the same real world, to be perceived through the same set of senses. A closer analogy to the human situation would result in GPS if all the task environments were basically analyzable by elementary programs in the general part of DPL. These would be programs for exploring the structure of objects and making tests of identity on the contents found in various places. Such a scheme not only seems more natural than the present one, but reinforces the vague feeling that there should exist a good set of “universal” differences.

REFERENCES

1. A. G. OETTINGER, Simple learning by a digital computer, *Proc. Assoc. for Computing Machinery*, September, 1952.
2. R. M. FRIEDBERG, A learning machine, Part I, *IBM Journal of Research and Development*, January, 1958.
3. A. NEWELL, J. C. SHAW and H. A. SIMON, The processes of creative thinking, P1320, The RAND Corporation, August, 1958.
4. A. NEWELL, J. C. SHAW and H. A. SIMON, Report on a general problem-solving program, P1584, January, 1959, The RAND Corporation.
5. A. NEWELL and H. A. SIMON, The logic theory machine, *Trans. on Information Theory*, vol. IT-2, No. 3, September, 1956.
6. O. G. SELFRIDGE, Pattern recognition and modern computers, *Proc. 1955 Western Joint Computer Conference*, Inst. Radio Engrs., March, 1955.
7. G. P. DINNEEN, Programming pattern recognition, *Proc. 1955 Western Joint Computer Conference*, Inst. Radio Engrs., March, 1955.

DISCUSSION

McCORMICK (*Naval Ordnance Laboratory, Corona, California*): What machine is used for your GPS work and how large is this program at the present time?

NEWELL: The program is on the RAND Johnniac. The work I am talking about here is all hand simulation. The GPS program is now being debugged, but is not yet solving any problems. Hence, most of our information about its characteristics comes from extensive hand simulation. The program is coded in an intermediate language called Information Processing Language IV, so it is a little hard to give you an idea of the size of the program. It totals about a thousand instructions in the interpretive language and uses the Johnniac 4000 words of core storage together with a 12,000 word drum.

KANTNER (*Armour Research Foundation, Chicago, Illinois*): Perhaps one can view the process as a discrete variable extremum process in that you apply an operator, select the operator based on an apparent difference to achieve presumably a smaller difference and the question that arises in my mind then is: what do you do about problems of local minima, that is, presumably you have the whole region over which you are operating full of trap spaces that are going to lead to problems?

NEWELL: This is taken care of first of all by an evaluation that recognizes when you are not making progress, when you are in a hole. Then there must be a way of getting out of it by having alternative methods of doing business. One way is for GPS to abstract the problem to a simpler space. In trying to solve the problem in this simplified space, it does different things than it did when it was brute forcing its way through the original space. In some of these methods GPS is also working forward just to see what it can do to progress a little bit. So you need a series of alternative methods to get you out of these traps.

MARKUS (*M.I.T.*): In your example, how much information do you have to provide GPS in order to get it to learn the differences?

NEWELL: This is an extremely hard question to answer without going into all the things in the paper I didn't present in the talk: that is, detailing the operators and the objects and so forth. The real gimmick, or the good trick as Don Campbell called it, is inventing the right spaces so that the operators we found and the differences we found made these tables of connectives small

enough so that there is not too much pointless search. Most of the information we put into the system in order to cause it to create those differences was introduced by constructing the right problem spaces for the system to work in. This I think was the big selection.

Now a whole wad of heuristics come in here. Suppose the system has two things it wants to differentiate. It applies an operator to them, then it looks at and compares the outputs. Now if the outputs are completely identical, it sees that no more processing will differentiate them. There is no sense in adding any more operators to the string in this programming language, so it quits at that point.

There is another principle that it uses: it is a good thing to use different concepts from the environment. This is a diversity principle. The problem is that there are a large number of possible basic things about the environment that it might attend to. All it knows is that there is a grab-bag full of possible concepts—fourteen or fifteen of them in the logic environment. Which one will it pick? It says, well, we need a principle of selection. Where do we find a principle of selection? It would be a good thing to pick those environmental features that are not used in the programs already constructed. This process will be laid out in the published paper fairly precisely and I can't give you a better or fuller answer to that now.

ANDERSON (*Burroughs*): In your description of GPS, you gave several examples of problems to be solved, all drawn from the field of mathematics. Could you suggest other problem types? To what extent is the selection of these problems to be solved influenced by the requirement of a formal system of transformation?

NEWELL: Let me start with the last part and work back. All of us in this game of building heuristic machines have tended to focus on highly formal systems like geometry, chess, checkers, systems of mathematical manipulation, because this was the kind of problem we thought we could deal with. Five years ago we didn't know how to deal with them. Now we can. I think the fact that these were formal systems has a great deal to do with our selection of them for the initial attempts.

On the issue as to whether it is essential that they be formal systems—this is now a personal answer—the answer is yes in one sense and no in another. Yes in the sense that by the time you get anything into a computer, it will look formal. No in the sense that I do not think you are going to find any system that is not formal in some meaning of the term.

This relates to the point made earlier about what happens if you try to program a fuzzy system like writing a novel. It is a very fuzzy problem, so to get started you rephrase it as a formal problem—and there is always a certain level at which it is formal, namely, a novel is written in letters spun together in sequence. So you can back off to where the system is formal, and get some purchase on it. When you do this, you have such a big space of possibilities, that you don't know how to write novels, so you add heuristics to see if you can get back to a smaller space. But you always find some way of formalizing.

On the issue of other problems for GPS, this again is highly opportunistic. We started on logic because we thought we could do it, and the GPS program was the successor to it. Trigonometry was a little program we worked out one weekend—a sort of gross heuristic for it—and it looked like a good next candidate for GPS.

We have tried GPS on chess. We now have a chess-playing program that is independent of GPS; so it was natural when we were thinking of how to make GPS more general, to say, well, we will play chess with it.

Well, the right answer to the question is that at this moment GPS will not play chess because we do not know how to make it play chess. That statement is equivalent to the fact that we have not been able to translate chess into a task of exactly the form required for GPS. By "exactly," I mean, so that you can really program it and it will really run on the computer and play chess. This gives me an opportunity to say this, to clear up just where GPS now stands: GPS could really be better called GPSWP—sort of GPS with pretensions. By "general" we don't mean that GPS can solve all problems. "General" means that it is built in such a way that the specific content of the problem area is factored from the general problem-solving heuristics. Therefore it is capable of tackling a wide class of problems. I don't know how big the class is. I can give you the characterizations; I can give you a specification that represents everything GPS needs to tackle a class of problems. Whether it will do a certain task effectively, or whether it can even work on it, depends on whether you can put the task in this form. That, at the moment, is a hard job in each individual case. Whether after you put the problem into this form, GPS will be able to solve problems well—whether there are enough heuristics to allow it to proceed—is a different question which we know even less about. So I can't specify what range of things GPS can do—we think of adding new ones all the time. A colleague, Lee Gregg, at Carnegie Tech, has been exploring a switch-throwing concept formation laboratory task and once every two months, some of us say to ourselves, "Why don't we try to make GPS do this task?" And we work on it for a while and it is too hard and then we drop it. So I can't say it will do this task.

LEARNING IN NEURAL SYSTEMS

P. M. MILNER*

*Department of Psychology,
McGill University, Montreal, Canada*

A FEW years ago stimulus-response conditioning models just about monopolized the field of psychological theory. To-day we have progressed to problem-solving models, a much more profitable approach, I think, but still not the complete story.

Behavior that is repeated goes through three main phases; the first is an exploratory phase in which the organism has not had any previous experience with anything resembling its present situation, so of course it cannot have any specific goal. Then there is the problem-solving phase when the organism *has* been in similar situations before, and *does* have a goal; and finally, after frequent repetitions of the behavior, the organism has the correct moves by rote. Some situations, like chess games, are too complex ever to be learned by rote, but many others, from tying a shoelace to reciting a poem or a geometry theorem, can be learned in that way. It is clear that these three phases overlap to some extent; a good psychological model has to be able to explain how they evolve, and how they are related to one another.

The reason I singled out the problem-solving phase as being most worthy of investigation is that the basic principles involved in the other two phases do not present any outstanding difficulty. There *are* plenty of difficulties, of course, it is just that they do not concern the underlying principles. As you know, it is possible to design a machine that will explore and make its way towards or away from a specified stimulus, if ever it comes within range. Devices such as guided missiles and Grey Walter's turtles are of this nature.

* This investigation was supported by a grant, M-2455, from the National Institute of Mental Health, U.S. Public Health Service.

Pure rote learning is even less of an achievement; a tape recorder is much better at it than any animal. The problem with rote-learning models, as formulated by psychologists, has been to discover why only useful, or at least selected, moves are learned in this way. The answer to that question is undoubtedly to be found in the preceding problem-solving phase, which is the main reason I shall spend most of my time on that phase to-day.

Before I start to construct models of any sort, though, I would like to spend a little time considering the fundamental neural processes that we must assume underlie all aspects of learning.

SYNAPTIC CHANGES IN LEARNING

It has frequently been pointed out that memory, in animals, appears to have two components which run different time courses. One of them decays quite rapidly, say within a minute or two, the other is for all practical purposes permanent. Reverberatory activity round nerve loops has been suggested as the mechanism for the first type of storage, but although the idea is theoretically feasible there are many practical difficulties, and some recent experiments initiated by Hebb seem to provide evidence against it.

Another hypothesis that has been proposed from time to time is that the stimulus produces an immediate large change of synaptic conductivity, most of which fades away fairly quickly, leaving a slight permanent residue. A feature of this type of trace is that it might easily result in reverberatory activity as a by-product; once a low resistance path has been marked out by the input signal, impulses might continue to circulate round closed parts of the path for some time. In this case the reverberation is not the primary mechanism of storage, it can only occupy the pathways opened up for it, but the reverberating impulses would keep the pathway open for a longer time, and they would probably increase the amount of residual "memory" resulting from a single exposure to a stimulus.

This hypothesis is attractive in many ways, but it suffers from the disadvantage that there is no satisfactory neurophysiological explanation for the synaptic change. The most difficult problem is why the change should occur only under certain conditions, i.e. only when the afferent synaptic knob and efferent cell body fire simultaneously. Some people, including myself at one time, have

tried to get round this difficulty by attributing the change to the whole of a neuron instead of to individual synapses. But this is very wasteful of potential storage, it divides the capacity by a factor of several hundred at least, and, the structure of the nervous system being what it is, it lands one with the additional problem of what to do about the hundreds of inappropriate connections that are acquired for every useful one.

Returning then to the individual synaptic change idea, neuro-biotaxis (the growth of active endings towards active cell bodies) produces the required effect but it would certainly take too long to account for the immediate component, and in any case, current knowledge about the microstructure of the nervous system almost entirely rules out the possibility of such a process. Another suggestion is that one or other of the apposed membranes at the synapse is somehow rendered more sensitive if they fire together, but not if they fire independently. Unfortunately this does not take into account the fact that the part of the cell membrane under the synaptic knob is depolarized every time the synaptic knob fires, irrespective of whether the rest of the efferent cell body follows suit. The biochemical and electrical effects of this inevitable sub-synaptic discharge are so much stronger locally than anything that the more distant parts of the cell can contribute, that it is hard to see what difference it could make to the synaptic change whether the cell body fired or not.

I am going to propose a mechanism that goes some way towards meeting these difficulties. It really doesn't involve anything more than the application of well-known electrophysiological principles. The idea is that in those parts of the nervous system where learning can take place some of the synaptic knobs are initially inaccessible to impulses arriving along the extremely thin terminal fibrils of the axon (i.e. the telodendra). Only when the efferent cell is on the point of firing is an impulse arriving along the telodendron able to get through to the synaptic knob. A virgin junction between telodendron and synaptic knob acts as a sort of gate which can only be opened by currents generated by the discharge of the efferent cell body. Thus the only time the knob fires is when the efferent cell is depolarized at the same time as an impulse arrives along the afferent axon. What lasting change takes place at the synapse after a knob has fired is anybody's guess, but at least there is a very profound

difference in the local conditions, depending on whether the knob has fired or not, to which the change can be attributed.

Figure 1 shows how I think the gating action works. As you may know, when a nerve impulse reaches a point where the membrane it is travelling along suddenly increases in diameter there is some

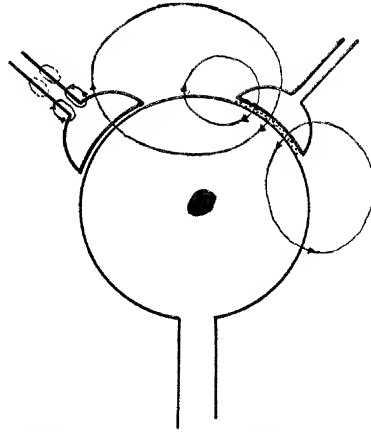


FIG. 1. The diagram illustrates a cell body with two of its synapses. The right-hand knob, which is not of the learning type, has just fired and the released transmitter substance is producing a heavy flow of depolarizing current through the sub-synaptic part of the cell membrane. This current is indicated by the large loops, one of which passes through the second, "learning" synaptic knob. The small loops of current shown at the junction of this knob with its telodendron are set up by an arriving afferent impulse. Normally the afferent impulse cannot provide the required current density to fire the membrane at the neck of the knob, where the area of the membrane suddenly increases, but when it is assisted by the current from the depolarized cell body the impulse gets past the critical point and the synaptic knob is discharged.

danger that it will die out. The current generated by the small surface area has to depolarize a much bigger area, and it may not be quite up to the job. This is especially likely in unmyelinated fibers because the current is not concentrated at nodal points but can spread through the whole membrane.

Blocking due to this effect has often been observed during experiments involving antidromic firing (i.e. firing a cell by electrical stimulation of the axon so that the impulse travels backward towards the cell). Eccles⁽¹⁾ gives examples of the antidromic impulse

failing to invade the cell body, and he even mentions that in the spinal cord, under some circumstances, impulses going in the normal direction may be blocked before reaching their synaptic knobs. The geometry at the junction between the telodendron and the synaptic knob is very similar to that at the junction of the axon and the cell body, on a much reduced scale, so that blocking at that point is by no means unlikely.

So, if we agree that the gate could normally be closed, how is it opened? The figure shows two synaptic knobs ending on the cell body; one of them, the one on the right, is not a "learning" synapse, it is always effective, and the cell has just been partly depolarized by an impulse reaching it. The depolarizing current, indicated by the large loops originating in the sub-synaptic region, has to leave the cell through the rest of the membrane, and as the cell body is encrusted with synaptic knobs some of it will flow through them also on its way back to the active region. Where the current flows into the knob at the synaptic surface it will hyperpolarize the membrane, but where it leaves, in the region of the junction with the telodendron, the current will depolarize the membrane. It will thus assist any current from an arriving impulse and allow it past the critical point so that it can invade the knob.* Once there, the impulse will release transmitter substance and produce a further depolarization of the efferent cell, and thus help to fire it.

An interesting corollary is that inhibitory afferents, which set up currents through the cell in the reverse direction, may block synapses that would normally conduct, thus augmenting the inhibitory effect by cutting off some of the cell's input.

If the hypothesis is correct it means that some interneural pathways have two switching points, one at the junction of the

* A rough calculation indicates that at the very most only 1% of the depolarizing current of the cell will pass through the membrane of the synaptic knob; 0.1% is a more likely estimate. The gate would have to be very delicately balanced to be influenced by such a small current except perhaps at the moment of complete depolarization of the cell. However, if the sub-synaptic membrane of the cell becomes permeable to potassium ions during the firing of the cell, the concentration of potassium below the synaptic knob could easily be increased by a factor of 10%, and that would produce a depolarization of several millivolts in the knob, quite enough to help an approaching impulse past the critical junction.

telodendron with the knob, and the other at the synapse itself. Once a knob has been invaded a few times, with whatever assistance is needed from the cell, it apparently becomes easier to fire, needing less assistance and finally none at all. If you want me to venture from what I consider to be fairly solid ground I might speculate that the gate stays open for a short time after the passage of an impulse because of the time it takes to repolarize the knob, or the cell membrane, or both. If another impulse comes along before the knob has completely recovered, it too would get past the junction, and thus keep the gate open for a further length of time, and so on. If there were a steady stream of afferent impulses it could keep the gate open for an appreciable period, seconds or minutes perhaps. Changes of a more permanent nature might be brought about by changes in the geometry or electrical conductivity of the surroundings, including the adjacent glial cells. A person of moderate ingenuity can think up a dozen other possibilities in as many minutes, a sure indication that we need more data.

I find this idea helpful because it means that the many desirable features of the synaptic change hypothesis that I mentioned earlier need not be ruled out on the grounds that they do not fit in with neurophysiological principles. A large immediate change of synaptic conductivity, followed by the required decay, is quite conceivable in terms of this mechanism. It is also always reassuring to find that you are dealing with processes that can be translated into electronic hardware; and, although by the time it was made to work a single "neuron" would probably take up as much room as the whole brain, I am sure one can be built to the foregoing specifications.

PROBLEM-SOLVING MODEL

The next step is to show how this neural mechanism can be used to store information in the problem-solving organism. This I shall try to do with the aid of Fig. 2. This figure is supposed to represent the nervous system, but the boxes are functional rather than anatomical units. However, in cases where it might make things clearer to some of you I shall say what part of the brain I think a box might represent (but don't take the localizations too literally). It will be obvious, I think, that these are not "black boxes"; what goes on in the boxes interests me very much. If it would not

make the diagram too messy I would be prepared to draw in the connections of sample neurons.

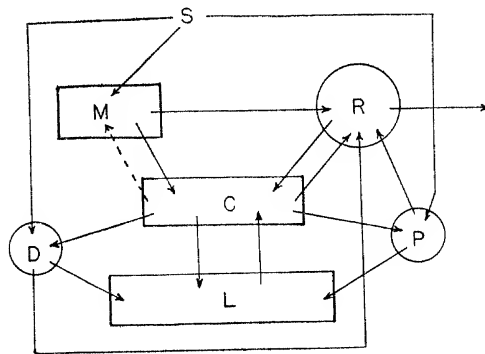


FIG. 2. Diagram of the connections of a decision-making learning system. (See text.)

The main information processing and storage networks are indicated by the rectangles *M*, *C*, and *L*. All are supposed to be in the cortex, and the paleocortex. *M* contains the cells that only fire when driven by some receptor activity *S*; they do not have enough feedback or closed loops to reverberate after the stimulus is removed. Some sorting takes place in *M*, as well as in the further networks, by means of converging-diverging overlap of connections. This ensures, by a process similar to cross-indexing, that some cells are uniquely fired by a particular stimulus pattern, an essential preliminary to any further processing of the information.

The next stage, *C*, is rich in internal connections; and the incoming connections from *M*, *R*, etc. are widely distributed throughout the area. Whereas the activity in *M* does not outlast the stimulus, the cells that are fired in *C* are able to keep firing one another for some time after the stimulus pattern has changed. Two main factors lead to the eventual decay of these reverberations; fatigue of the synapses whose resistance was reduced by the stimulus, and interference from succeeding patterns of input, which have inhibitory as well as excitatory components. Miller⁽²⁾ has pointed out that there is some reason to suppose that the human adult can sustain about seven independent traces at one time in this reverberatory state.

Presumably if the number exceeds this the mutual interference gets too severe.

The third part of the system, *L*, is an additional storage space intended to explain the increased learning that goes on under conditions of reinforcement, reward or punishment. There are few direct interconnections from one cell to another within the system, but it is freely connected in both directions to the cells of *C*, so that there is ample opportunity for indirect connections between *L* cells via *C* cells.

Now let us turn to the primitive organism represented by the units in the circles. This is the goal-seeking automaton I spoke of at the beginning. It can explore, or move about at random if that deserves the name of exploration, and when stimuli from some needed commodity come within range of its receptors it can descend upon the source and consume it. The organism is a sort of optional-goal guided missile. The part *D* represents the reward or positive drive system; a complex of nuclei many of which seem to be located in and around the anterior hypothalamus. Some of its cells are sensitized by deprivation of various sorts, some by sex hormones, carbon dioxide tension, and so on. A few of these cells may fire, or increase their rate of spontaneous activity in the sensitized condition, but the main effect is not to fire the cell but to lower its threshold to stimuli from an appropriate satisfier. Such an arrangement must be an evolutionary development; an animal that was attracted to water when hungry and went to sleep when short of water would most likely never wake up again to perpetuate its line.

The punishment system, *P*, could perhaps be considered as part of the drive system having a reciprocal relation to the rest. I have shown it as separate from *D* because it is more convenient that way if you want to talk about approach-avoidance conflict situations, i.e. those in which a stimulus pattern has acquired associations with both reward and punishment. Another difference between *P* and at least some parts of *D* is that *P* is always fired by the appropriate stimuli without needing any preparation; no hormones or pain deprivation are needed to sensitize the punishment system.

Both these drive systems are connected to the motor system *R*. This I consider to be an extensive and complicated system stretching

from the motor cortex through the striatum and subthalamic region to the medulla. It is capable of spontaneously generating activity which, in conjunction with feedback from the proprioceptive system, produces basic acts such as walking, grasping, chewing, sucking, breathing, and so on. The innate connections may provide only crude versions of these acts which can later be improved by learning.

If this is beginning to sound a bit vitalistic I would point out that there is nothing at all mysterious about a system going into complicated multiple oscillations if it is provided with energy from some outside source. The fact that the neural oscillations of the system have an innate tendency to produce beneficial rather than self-destructive behavior is another achievement of natural selection, closely related to what I just said about the drive system.

Now, to see how the bits of our simple automaton work together, suppose the random activity of R in a hungry animal brings it into the field of a piece of food. Immediately the activity in D increases, and this is reflected in an increased firing of R ; the animal continues to do whatever it was doing but more energetically. As long as the stimuli from the food keep increasing in intensity the animal will keep going, but as soon as they decrease, that particular activity of R will be attenuated and may stop. (Adaptation and fatigue in neural systems differentiate the transmitted signal with respect to time like a series capacitance in an electronic circuit, so the activity in R may follow the *change* of intensity of the food stimuli rather than their absolute value.)

As soon as the ongoing activity of R falls to a low value another mode of activity will take its place, and if this new activity produces a random movement which increases the stimuli received from the food again, it will increase the feedback from D to R and that response will be maintained as long as it works. In this way anything the animal does which brings it towards the food (or makes the food stimuli stronger) will be vigorously pursued, and anything which reduces the stimuli will be abandoned. Such behavior would almost always result in eventual contact with the food, at which point other reflexes take over. If the animal runs into any trouble on the way, the P system will be fired and it will immediately stop the current R activity, replacing it with a withdrawal reflex or other defensive reaction.

So the mechanism represented by the outside track of the diagram, through *S*, *D*, *P*, and *R* can bring the animal to what it needs and keep it out of trouble; but only very inefficiently, and only if the environment is well stocked with the necessities of life. If random exploration will not bring it within smelling distance of food before it is too weak to search further, the simple animal has had it. Our learning boxes, *M*, *C*, and *L* can be considered to provide an alternative feedback path within the organism which can function when the feedback path through the environment is broken, i.e. when the goal is out of range of the animal's receptors. This alternative path is the one from *R* through *C* to *D* and back to *R* again. Of course, it doesn't do any good unless the right information is stored in the networks.

Very briefly what happens if all goes well is that some activity that could lead to a response starts up in *R* and is fed into *C* where it interacts with the information stored there, and with the current input, *S*. If this combination in *C* evokes associations with *D*, or *P*, it will stir up activity in those boxes. Any *R* activity that increases the firing in *D* in this way has a much better chance of building up into an overt response than others (just as it would have been intensified if it had increased the firing in *D* by bringing the animal physically nearer to food).

Now we come to the real problem of learning, how the connections are established in the various networks by experience. I follow Hebb in assuming that such learning occurs in two stages, an initial stage of building up cell assemblies (i.e. patterns of neural connections that represent frequently occurring stimulus and response patterns), followed by a second stage in which the cell assemblies acquire connections with one another in accordance with sequences of environmental events.

I have already stated that when a stimulus is presented it fires cells in *M*, which stop as soon as the stimulus is withdrawn, and in *C*, where the activity persists by means of connections between the affected cells. These interconnections have synapses of the "learning" variety, so that paths which were unavailable before the stimulus was presented can afterwards conduct impulses freely for a time. Thus a group of cells fired by a stimulus can keep itself firing in a reverberatory manner, and the synapses involved become a little more easily traversed ever afterwards. If the same stimulus

(or others similar enough to fire a lot of cells in common with it) is presented repeatedly, the internal connections of the group of cells it fires will become permanently effective; the interlinked group is then a cell assembly. This does not mean that reverberations continue indefinitely in the assembly, it merely means that a stimulus that only reaches a fraction of the cells in the group will henceforth be able to trigger off the complete reverberation. In other words, just a part of the original stimulus would be an adequate stimulus after a cell assembly is formed; or, more important from the point of view of further learning, if a number of the cells of one assembly acquire connections from another assembly, the reverberation of the second might be adequate to trigger a reverberation of the whole of the first. It is not difficult to see how such interconnections between assemblies could be acquired. Any learning synapses from cells in one assembly on to cells in another will have a finite probability of being opened up if the two assemblies are reverberating at the same time.

I am afraid I now only have time to give a compressed and sketchy example of the way specific associations are acquired, and then used in later performances. Exploration can be considered as a sequence of R 's interspersed with changes of S . S_0 followed by R_1 leads to S_1 ; then the next response R_2 leads to S_2 , etc. The R 's occur at random as a result of the operations of the motor system, the S 's are contingent on the environment. The reverberations that build up in C during all this soon get very complicated, but at the start the cell assembly S_0 will acquire connections with the activity fed into C from the activity R_1 , and the pair of them will acquire connections with the assembly for S_1 when it appears. Eventually the animal reaches a stimulus that fires D , and the automatic goal-seeking mechanism takes over. When D fires it brings in a lot of cells in L that were not firing before, and provides a great many more opportunities for links between any reverberations there may be in C . These reverberations also acquire connections to the cells fired in D by the satisfying stimulus.

Later, if the animal ever finds itself in a situation where the deprivation is the same as before (i.e. the same D cells are sensitized) and the stimulus pattern arouses one of the cell assemblies that was reverberating when the satisfying stimulus appeared, that cell assembly will increase the firing of D . If the activity in R now

happens to adopt the pattern it had on the previous occasion it will activate more assemblies in *C* that have connections with *D* and so produce a further increase; then the feedback from *D* to *R* will raise the activity there to the level of an overt response. Other modes of *R* activity may decrease the activity in *D* by interfering with the reverberations in *C* that have associations with *D*. Thus, *on the average*, any mode of activity that *R* adopts that is different from the one that previously led to the reward will be less likely to become an overt response. The bias in favor of the previously successful response may be slight but it must eventually pay off in an increased frequency of that response.

Waiting for the right *R* activity to come along is very time consuming and inefficient. It would be better if the *R* system could be joggled into producing the right activity straight off when the appropriate stimulus came along.* The connections that do this must not be acquired too readily or we shall run into the danger of witless stereotypy that we hoped to avoid by abandoning the simple *S-R* formulation. However, as you can see, the problem-solving phase generates a higher than chance frequency of correct responses, so that if there is *slow S-R* learning on a frequency basis (via the path shown from *C* to *R* on the slide) some of the cells involved in the correct response activity will eventually get connected to the cell assembly for the stimulus that most often precedes it. Then the animal will no longer have to wait for long periods at a choice point until its *R* box grinds out an acceptable pattern of response activity (or, as we are apt to say, it makes up its mind what to do); the stimulus will bias the *R* system in the direction of the most frequently made response.

Notice, however, that when the activity is started in *R* by this means it still has to be passed by the *C*, *D* and *P* circuits. The stimulus, as it were, suggests the response that should be given priority of consideration, but if the animal is not hungry, or got a bad shock last time it made that response, the "suggested" *R* activity may still not become an overt response.

*I was reminded at the Conference that the desirability of this feature was first suggested to me three or four years ago by Dr. Minsky. As I remember, neither of us was able to fit it into the model I was grappling with at the time.

SUMMARY

To sum up, this relatively simple model can thus explain in a rudimentary way the three phases of learning that I enumerated at the beginning; exploration, problem-solving, and rote learning. The pure exploration and rote-learning phases require no elaborate explanations; both can be reproduced by machines operating on well-known principles, and there is enough evidence to make it pretty clear that the nervous system could employ similar principles. The problem-solving phase can best be summarized as vicarious exploration; response tendencies being sampled until one comes along that increases the associations with the reward system (or decreases the associations with the punishment system in avoidance training) and this response is then emitted.

I based the model on a neural mechanism of learning in which some specialized "learning" synapses are initially completely ineffective, but can be switched to complete effectiveness if the cell body on which they end is strongly depolarized. If an impulse arrives at the synapse at such a time it traverses it and leaves it in a completely conducting state for a short time. Each time the synapse conducts it stays in the conducting state longer, and needs somewhat less depolarization of the efferent cell body to open it up again after the transient conducting state has decayed.

It is my hope, of course, that the model I have described uses the same general principles as the brain; but even if it does not I think it should still work. If it were practical to assemble the enormous storage capacity (about 10^{12} bits I believe in the human), and spend enough time adjusting the numerous parameters, I think it would be possible to construct a rational machine along the lines of the model.

REFERENCES

1. J. C. ECCLES, *The Neurophysiological Basis of Mind*, p. 124, Clarendon Press, Oxford, 1953.
2. G. A. MILLER, The magical number seven, plus or minus two: some limits on our capacity for processing information, *Psychol. Rev.*, **63**, 81-97 (1956).

DISCUSSION

DAWE (*Office of Naval Research*): I would like to raise the question of the usefulness of the reverberating circuit as the mechanism for memory traces. The reason I raise this question is that Lyman, a number of years ago, and

Stonehauser more recently, has shown in hibernating mammals the electroencephalogram disappeared entirely in hibernating animals. Doesn't this suggest that memory may be a chemical change?

MILNER: Yes, I think I explained that the reverberatory trace, in as much as I am using it, it is a very short-term process, and it may be that this holds the information long enough for a chemical or other change to take place. I think I mentioned that Hebb has done some experiments (to be reported at the Montevideo symposium on Brain Mechanisms and Learning in August) which show that you never get a pure reverberatory trace. It always leaves some residue.

ROSENBLATT: I would like to pose a question for comment. Fairly early in your discussion you indicated that you considered it essential that there be sets of cells uniquely fired by each possible stimulus that might be presented. There seems to be very little biological evidence that cells exist which are uniquely fired in this fashion. Moreover, we have several models now, including my own, in which this is shown not to be an essential feature of the system.

I would not be so critical on this particular point except it seems to me that the inferences about the behavior of the cell assembly with respect to being fired by parts of adequate stimuli after the thresholds are lowered falls very much in its logical characteristics on this assumption that the cell assembly in the first place derives from cells that respond to particular stimuli. If the threshold is going down progressively, admittedly part stimuli do become sufficient to fire the cell assembly, but at the same time any new stimulus which embodies these parts become sufficient and the population of sufficient stimuli increases in a geometric fashion.

It seems to me at the same time you are making the cell assembly susceptible to firing by part stimuli, you are simply broadening it out so if it is fired by a much wider population of stimuli than could fire it in the first place, that most members of this population will be quite irrelevant as against the originally sufficient stimuli.

MILNER: I was really not proposing to discuss the problem of perceptual learning to-day but I do have some ideas about this that are not quite the same as yours, I think. Shall we leave it until later?

NEWELL: There seems to be two things going on here. One is a set of notions about basic building blocks, if I can use that word, and the other is to build it up to a functional organization that we hope will solve our problems. I wasn't able to follow it close enough to answer my own question and I wonder if you have done some thinking along the line of what alternative realizations there could be at the functional level so one could test the adequacy—not worrying about the building blocks at the moment, but even assuming in some sense you could just store some information the way you were talking about, but still going to radically different representatives of this so you could see the functional adequacy of this kind of a model, say at the level of the GPS or the programs we were talking about.

MILNER: Unfortunately, I know really nothing about computers or programming so I am at a disadvantage when it comes to simulation of that sort. I have worked this out much more elaborately just on paper, the question of what would happen in a very simple situation, and under these simple conditions. It seems to work as far as I can see, semi-intuitively. I wouldn't mind if somebody tried to simulate this more formally in a cleaner way, but I haven't done it and I am not sure that I could.

ARMSTRONG (*IBM, Wappingers Falls, New York*): I wonder if you would say something about the kind of sorting you are doing in box *M*?

MILNER: Well, that I think is probably one of the problems that might be answered better this afternoon if we have time. I think it is very similar to the sorting that Uttley has suggested and explained in his paper in Automata Studies. Convergent-divergent systems of neural connections can act like a sort of cross-indexing, selecting neurons at a central level which respond only to one particular stimulus pattern.

BLIND VARIATION AND SELECTIVE SURVIVAL AS A GENERAL STRATEGY IN KNOWLEDGE-PROCESSES

DONALD T. CAMPBELL

*Department of Psychology,
Northwestern University, Evanston, Illinois*

IN AN august group such as this—august not with age and past distinction—august rather with youthful vigor, mathematical sophistication, and status as the creators of a new and esoteric discipline—in an august group such as this, I can only speak with comfort by first avowing my own position as an outsider. I am an old-fashioned psychologist, still preoccupied with the chronic problems of psychology, problems that first separated psychology from philosophy, that opposed E. L. Thorndike and William McDougall in 1900, and that have more recently opposed behaviorism and gestalt psychology. The source of my presumption in being willing to address you at all comes from my conviction that from your hands, that is, from servomechanism theory and machine simulation of mental processes, are coming the most promising resolutions of these traditional problems.

The particular facet of psychology which I wish to represent today can be designated “the comparative psychology of knowledge-processes.” This activity was first prominent prior to the clean break between psychology and philosophy, as epitomized in the work of James Mark Baldwin. The subsequent departmental separation of the two fields was accompanied by an almost total cessation of such endeavour. Philosophers were left in temporary monopoly of all aspects of the problem of knowledge, but on their part came to reject the greater part of the task: logicians combatted

"psychologism" in logic along with combatting the once concomitant faith in a single correct logic which would describe man's thought processes. The analytic philosophers have by and large tended to reject as a proper philosophical task the whole problem of induction, once satisfied that they could not solve it deductively. Into the area of neglect thus created, psychologists have recently begun to move again. It is no accident that the pioneers in this direction, such as Piaget (1950) and Lorenz (1943, 1951), have been men who have also recognized psychology's need for the contributions of cybernetics.

My own efforts along this line have employed the general tactic of comparing knowledge processes at varying levels, looking both for parallels in process and for differences (Campbell, 1959). For example, monocular visual cues have been employed in one instance to suggest operational indices of the degree to which an aggregate of persons constitute a social entity (Campbell, 1958) and, in another instance, binocular parallax has been used to suggest a criterion for the convergent and discriminant validity of psychological tests (Campbell and Fiske, 1959, as interpreted in Campbell, 1959). My theme to-day deals with a less romantic source of parallels than vision provides, a simpler process which is much more fundamental and primitive. And while my remarks are perhaps most appropriately addressed to my fellow psychologists, there is one sense in which they may be appropriate here: the major inspiration for my theme of to-day has come from your ilk, and presenting these remarks here constitutes something on the order of a feed-back, enabling you to sample your influence in the underdeveloped areas of science, and to estimate the noise of the channel when the communication of your message has been dependent upon *words*, unsupported by either mathematics or hardware.

As used here, the genus *knowledge processes* includes both self-organizing systems and systems whose adaptive reorganization is achieved by other means. The higher knowledge processes represent self-organizing systems. On the other hand, in the category of self-organizing systems there is included more than knowledge processes. But all self-organizing systems in which the reorganization proceeds in the direction of a better fit to an external environment would be included as knowledge processes. Within the arena thus delimited, this paper will work toward these general conclusions:

1. A blind-variation-and-selective-survival process is fundamental to all inductive achievements, to all genuine increases in knowledge, to all increases in fit of system to environment.
2. The processes which shortcut the full blind-variation-and-selective-survival process are in themselves inductive achievements containing wisdom about the environment achieved originally by a blind-variation-and-selective-survival process.
3. In addition, such substitute processes contain in their own operation a blind-variation-and-selective-survival process at some level.

Were this a group of psychologists or philosophers, this dogmatic pronouncement could be counted on to provoke some disagreement, particularly by the time it has been extended to encompass creative intellectual achievements. Even among you, my acceptance and extension of the position outlined by Ashby in his *Design for a Brain* must make me responsible for the practical inadequacies of his solution.

Between a modern experimental physicist and some virus-type ancestor there has been a tremendous gain in knowledge about the environment. In bulk, this has represented inductive achievements, expansions of knowledge beyond what could be deductively derived from what was already known, "break-outs" from the limits of available wisdom. If such expansions had represented wise anticipations, they would have been exploiting full or partial knowledge already achieved. Thus the real gains must have been the products of explorations lacking foresight or prescience, and in this sense blind, or stupid. The successful explorations were in origin as blind as those which failed, and the difference between them due to the nature of the environment, the difference between them, once encountered, representing a gain in wisdom about the environment.

The general model for such inductive gains is that underlying trial-and-error problem solving and natural selection in evolution (Baldwin, 1900; Ashby, 1952; Pringle, 1951). Three conditions are necessary: a mechanism for introducing variation, a consistent selection process, and a mechanism for preserving and reproducing

the selected variations. In what follows we shall look for these three ingredients at a variety of levels. But first a comment on the use of the word "blind" rather than the more usual "random." It seems likely that Ashby (1952) unnecessarily limited the generality of his mechanism in Homeostat by an irrelevant effort fully to represent all of the modern connotations of "random." Equiprobability is not needed and is strongly abrogated in the mutations which lay the variation base for organic evolution. Statistical independence between one variation and the next, while frequently desirable, can also be abrogated. In particular, for the generalizations essayed here, certain processes involving systematic sweep scanning are recognized as *blind*, in so far as variations are produced without prior knowledge of which ones, if any, will furnish a selectworthy encounter. A first important connotation of blind is that the variations emitted be independent of the environmental conditions of the occasion of their occurrence. While this too can be abrogated in considerable degree, if the correlation is high it reduces the possibility of novel adaptations. A second important connotation is that the occurrence of trials individually be uncorrelated with the solution, in that specific correct trials are no more likely to occur at any one point in a series of trials than another, nor than specific incorrect trials. In so far as observation of self-organizing systems shows this to be abrogated, the system is making use of already achieved knowledge, perhaps of a general sort. The prepotent responses of an animal in a new puzzle box may thus represent prior general knowledge, transferred from previous learning or inherited as a product of the mutation and selective survival process. A third essential connotation of *blind* is rejection of the notion that a variation subsequent to an incorrect trial is a "correction" of the previous trial or makes use of the direction of error of the previous one.*

In this perspective, the epistemologically most fundamental knowledge processes are embodied in those several inventions making possible organic evolution. At the already advanced level of cellular life, this is a "learning" on the part of the species by

* In so far as mechanisms do seem to operate in this fashion, there must be operating a substitute process carrying on the blind search at another level, or feedback circuits selecting "partially" adequate variations, providing information to the effect that "you're getting warm" etc.

the blind variation and selective survival of mutant individuals. In terms of the three requirements, variation is provided by the mutations, selection by the somewhat consistent or "knowable" vagaries of the environment, and preservation and duplication by the complex and rigid order of chromosome mitosis. Bisexuality, heterozygosity, and meiotic cell division represent a secondary invention increasing the efficiency of the process through increasing the range of variation and the rate of readjustment to environmental shifts. The selection and preservation processes remain the same. The ubiquity of bisexuality, its several independent inventions, and the multifarious elaboration of the theme, all speak to its tremendous usefulness. If we were to consider the species or the gene-pool of the breeding group as the "system," then these are self-organizing processes.

If our focus to-day is upon more sharply bounded entities, we must look to higher evolutionary developments which shift a part of the locus of adaptation over to processes occurring within the single organism. Numerous such processes exist, each not only a device for obtaining knowledge, but also representing general wisdom about environmental contingencies achieved through organic evolution, making possible more efficient achievement of "local" knowledge. One of the most primitive of these is exploratory locomotion, described in the protozoa by Jennings (1906) and accepted as a model for Homeostat by Ashby (1952). Forward locomotion persists until blocked, at which point direction of locomotion is varied blindly until unblocked forward locomotion is again possible. The external physical environment is the selection agency, the preservation of discovery is embodied in the preservation of the unblocked forward movement. At this level, the organism has "discovered" that the environment is discontinuous, consisting of penetrable regions and impenetrable ones, and that impenetrability is to some extent a stable characteristic—it has learned that it is a better strategy to try to go around than to wait until one can move through.

In so far as animals without distance receptors, such as paramecia and earthworms, can learn through contiguity, the species has already learned that there is some event-contingency stability in the environment. In the degree to which such processes are useful, there is a discovery of slower transformation processes on the part

of relevant segments of the environment than of the organism. In addition, whereas the ultimate selection is life or death in encounters with the external environment, by the evolutionary stage at which learning is possible, much of this once-external criterion has been internalized. Crude environmental contingencies with low selection ratios are now represented as pleasures or pains, or as reinforcers more generally. The selection becomes much more sharp, but the contact with the environmental realities less direct. The presence of a fundamental trial-and-error process in learning needs no elaboration or defense. Suffice it to say that recognition of such a process is found in all learning theories which make any pretense of completeness, including at least three of Gestalt inspiration (Campbell, 1956*a*). While higher vertebrate (and higher cephalopod) learning makes far more use of the short-circuiting of overt trial-and-error by vision than is allowed for by the usual behavioristic learning theory (Campbell, 1956*b*), for convenience here the multiplication of levels will be avoided by treating trial-and-error learning as a single process-level.

The next and most striking class of discoveries are those centering around echo-location and vision. Pumphrey (1950) interprets the primitive sense-receptor of the fishes called the "lateral-line organ" as a crude echo-location device, making use of the reflected pulses of the fish's own swimming. Griffen (1958) has documented in detail the use by bats and cave birds of sonic and supersonic vocalizations selectively reflected by obstacles of the environment. Kellogg has made a similar case for porpoises (1958). Here is a powerful substitute for blind locomotor exploration. (See Simon, 1957, p. 264, for an estimate of such gains.) A wave pulse is emitted blindly in all directions. The obstacles of the environment selectively reflect from certain of these directions, and thus provide a feedback substitutable for that which would have been received had the animal locomoted in that direction. Radar guidance systems employ an analogous substitution of a blindly scanning electromagnetic wave pulse, in economical substitution for a blind scanning of the same environment of potential locomotions by full ship or projectile movements.

Visual perception seems interpretable as a substitute search process of similar order. The full analogy is weakened by the absence of an emitting process on the part of the organism. Instead,

advantage is taken of diffuse electromagnetic waves made available from external sources. Consider first a pseudo-eye consisting of but a single photo receptor cell. (Such a device has been distributed for use by the blind in which a photocell output is transformed into a variable pitch sound.) With such a device, blind scanning as in a radar system is essential. Brightness contours can be located and fixated by continual crossing, as in a "hunting" control process. To conceive of such an eye as a blind searching device substituting for a more costly blind locomotion in the explored directions is not difficult. The eyes of insects and vertebrates and the higher cephalopods differ from such a device by having multiple photocells, making possible selective reflection from objects in multiple directions at once. Each receptor cell can be conceived of as exploring the possibilities of locomotion in a given direction, the retina collectively as thus exploring the possibilities of locomotion in a wide segment of potential directions for locomotion. Except as the eye is aimed by other sources of knowledge, these possibilities have been made "blindly" available without prescience or insight (Campbell, 1956b). For the "blindness" of an eyeless animal there has been substituted a process so efficient that we use it naively as a model for "direct," unmediated knowing. But the process is still one of blind search and selective survival, in the sense employed in this paper.

Vision is a very complex and marvelous mechanism, and the brief presentation here does not do justice even to the random search components involved. Hebb (1949) has well documented the active search of eye movements, correcting the model of the inactive fixed-focus eye which is implicit in both Gestalt psychology and conditioning theory. Riggs (Riggs, Ratliff, Cornsweet and Cornsweet, 1953; Riggs, Armington and Ratliff, 1954) and Ditchburn (Ditchburn, 1955; Ditchburn and Fender, 1955) have documented the essential role of the continuous low amplitude scanning provided by "physiological nystagmus" or "fixation tremor." Platt (1958) has provided a brilliant analysis of the role of a blind "rubbing" process, his "lens-grinding" model for the achievement of visual acuity and spatial representation in a visual system containing unaddressed elements. These and other considerations convince me that although vision represents the strongest challenge to the generality of a blind-variation-and-selective-survival process, it is

not in fact an exception. These brief comments have not fully justified this conclusion, however.*

Taking these echo-location and visual exploratory processes collectively, several general aspects can be noted: all exploit a specific and limited coincidence, i.e. that objects impenetrable by organismic locomotion also are opaque to, or reflect, certain wave forms in the acoustical and adjacent frequencies and in the bands of electromagnetic waves of the visual and radar spectra. It is this coincidence, unpredictable upon the basis of the prior knowledge available to the more primitive organisms, which makes possible such marvelously efficient shortcuts. While phenomenologically vision is more "direct" than other knowledge processes, it is seen in this perspective as an indirect, substitute process. As in all substitute knowledge processes, the effectiveness is limited by the accuracy of the coding process, i.e. the translation terms between one level and another. Such coding is never exhaustive (Platt, 1956). It always involves abstraction, and along with this some fringe imperfection and proneness to systematic error (Campbell, 1958*b*). It must finally be checked out and corrected by overt locomotion. Its efficacy is limited by the *relevance* of the coding to the more fundamental level of behavior for which it is a substitute. This relevance was itself initially tested out by a blind-variation-and-selective-survival process at the level of organic evolution or early childhood learning. (Species differ in this regard.) The phenomenal directness of vision tempts us to make vision prototypic for knowing at all levels, and leads to that chronic belief in the existence of "direct" and "insightful" mental processes which it is a major purpose of this paper to deny.

CREATIVE THOUGHT

Creative thought provides the next level knowledge-process for the present discussion. At this level there is a *substitute* exploration of a *substitute* representation of the environment, the "solution" being selected from the multifarious exploratory thought-trials according to a criterion *substituting* for an external state of affairs. In so far as the three substitutions are accurate, the solutions when

* Thurstone (1924) provides an undetailed interpretation of perception as a substitute trial-and-error process, in a book which in many ways anticipates the contributions of cybernetics to psychology.

put into overt locomotion are adaptive, leading to intelligent behavior which lacks overt blind floundering, and is thus a knowledge-process. To include this process in the general plan of blind variation and selective survival, it must be emphasized that the internal emitting of thought-trials one-by-one is "blind," lacking prescience or foresight. The process *as a whole* of course provides "foresight" for the overt level of behavior, once the process has blindly stumbled into a thought-trial that "fits" the selection criterion, accompanied no doubt by the "something clicked," "Eureka" or "aha-erlebnis" that marks the successful termination of the process.

To-day, we find the blind-variation-and-selective-survival model most plausibly applied at the levels of organic evolution and trial-and-error learning of animals. Historically, however, the phrase "trial-and-error" was first used to describe thinking by Alexander Bain as early as 1855, two years before Darwin's publication of the doctrine of natural selection. Not only for historical interest, but also to further develop the psychology of creativity, the following quotations from him are provided:

"Possessing thus the material of the construction and a clear sense of the fitness or unfitness of each new tentative, the operator proceeds to ply the third requisite of constructiveness—trial and error— . . . to attain the desired result. . . . The number of trials necessary to arrive at a new construction is commonly so great that without something of an affection or fascination for the subject one grows weary of the task. This is the *emotional* condition of originality of mind in any department." (Bain, 1874, p. 593.)

"In the process of Deduction. . . . The same constructive process has often to be introduced. The mind being prepared beforehand with the principles most likely for the purpose . . . incubates in patient thought over the problem, trying and rejecting, until at last the proper elements come together in the view, and fall into their places in a fitting combination." (Bain, 1874, p. 594.)

"With reference to originality in all departments, whether science, practice, or fine art, there is a point of character that deserves notice. . . . I mean an Active turn, or a profuseness of energy, put forth in trials of all kinds on the chance of making

lucky hits. . . . Nothing less than a fanaticism of experimentation could have given birth to some of our grandest practical combinations. The great discovery of Daguerre, for example, could not have been regularly worked out by any systematic and orderly research; there was no way but to stumble upon it. . . . The discovery is unaccountable, until we learn that the author . . . got deeply involved in trials and operations far removed from the beaten paths of inquiry." (Bain, 1874, p. 595.)

Ernst Mach was another great nineteenth century thinker about thinking who emphasized this model. We here to-day remember him most as a psychologist-physicist-philosopher who contributed to the present-day positivistic recognition of the hypothetic character of our constructions of the world and who made clear to young Einstein the empirical presumptions involved in physicist's assumptions of a Euclidian space. But when, at the age of 57, in 1895, he was called back to his *alma mater* the University of Vienna to assume a newly created Professorship of the History and Theory of Inductive Science, he chose a quite different theme for his inaugural address. His title was "On the part played by accident in invention and discovery." The occasion indicates the importance he gave to the message, and indeed, his paper is a neglected classic in the psychology of knowledge processes. These quotations will further reinforce the model of creative thought being presented, and make it clear that it was available before the days of machine simulation of thinking:

"The disclosure of new provinces of facts before unknown, can only be brought about by accidental circumstances. . . ." (Mach, 1896, p. 168.)

"In such [other] cases it is a psychical accident, to which the person owes his discovery—a discovery which is here made "deductively" by means of mental copies of the world, instead of experimentally." (Mach, 1896, p. 171.)

"After the repeated survey of a field has afforded opportunity for the interposition of advantageous accidents, has rendered all the traits that suit with the word or the dominant thought more vivid, and has gradually relegated to the background all things that are inappropriate, making their future appearance impossible; then from the teeming, swelling host of fancies which a free and high-flown imagination calls forth, suddenly that

particular form arises to the light which harmonizes perfectly with the ruling idea, mood, or design. Then it is that that which has resulted slowly as the result of a gradual selection, appears as if it were the outcome of a deliberate act of creation. Thus are to be explained the statements of Newton, Mozart, Richard Wagner, and others, when they say that thoughts, melodies, and harmonies had poured in upon them, and that they had simply retained the right ones." (Mach, 1896, p. 174.)

Poincaré (1908, 1913) in his famous essay on mathematical invention presents a point of view which is also judged to be in agreement (although Hadamard, 1945, has not interpreted it as so). He first gives an example in imagery:

"One evening, contrary to my custom, I drank black coffee and could not sleep. Ideas rose in crowds; I felt them collide until pairs interlocked, so to speak, making a stable combination." (1913, p. 387.)

Poincaré feels that it is rare for this blind permuting process to rise into conscious awareness, and that as a rule only the successful selected alternatives enter consciousness. The level of consciousness involved is of course not crucial to the model here presented. The restraints on complete randomness imposed by Poincaré are acceptable under the general model as representing the application of prior knowledge. Because of the relevance of Poincaré's comments, and because Hadamard (1945) has cited him in opposition to the accidentalist position while he is read here as favoring the selective-survival version of it, these longish excerpts are read into the record:

"It is certain that the combinations which present themselves to the mind in a sort of sudden illumination, after an unconscious working somewhat prolonged, are generally useful and fertile combinations, which seem the result of a first impression. Does it follow that the subliminal self, having divined by a delicate intuition that these combinations would be useful, has formed only these, or has it rather formed many others which were lacking in interest and have remained unconscious?

"In this . . . way of looking at it, all the combinations would be formed in consequence of the automatism of the subliminal self, but only the interesting ones would break into the domain of consciousness. And this is still very mysterious. What is the

cause that, among the thousand products of our unconscious activity, some are called to pass the threshold, while others remain below? Is it a simple chance which confers this privilege? Evidently not; among all the stimuli of our senses, for example, only the most intense fix our attention, unless it has been drawn to them by other causes. More generally the privileged unconscious phenomena, those susceptible of becoming conscious, are those which, directly or indirectly, affect most profoundly our emotional sensibility." (1913, p. 391.)

"... we reach the following conclusion: The useful combinations are precisely the most beautiful, I mean those best able to charm this special sensibility that all mathematicians know, but of which the profane are so ignorant as often to be tempted to smile at it.

"What happens then? Among the great numbers of combinations blindly formed by the subliminal self, almost all are without interest and without utility; but just for that reason they are also without effect upon the esthetic sensibility. Consciousness will never know them; only certain ones are harmonious, and, consequently, at once useful and beautiful. They will be capable of touching this special sensibility of the geometer of which I have just spoken, and which, once aroused, will call our attention to them, and thus give them occasion to become conscious.

"This is only a hypothesis, and yet here is an observation which may confirm it: when a sudden illumination seizes upon the mind of the mathematician, it usually happens that it does not deceive him, but it also sometimes happens, as I have said, that it does not stand the test of verification; well, we almost always notice that this false idea, had it been true, would have gratified our natural feeling for mathematical elegance.

"Thus it is this special esthetic sensibility which plays the role of the delicate sieve of which I spoke, and that sufficiently explains why the one lacking it will never be a real creator.

"Yet all the difficulties have not disappeared. The conscious self is narrowly limited, and as for the subliminal self we know not its limitations, and this is why we are not too reluctant in supposing that it has been able in a short time to make more different combinations than the whole life of a conscious being could encompass. Yet these limitations exist. Is it likely that it is able to form all

the possible combinations, whose number would frighten the imagination? Nevertheless that would seem necessary, because if it produces only a small part of these combinations, and if it makes them at random, there would be small chance that the *good*, the one we should choose, would be found among them.

"Perhaps we ought to seek the explanation in that preliminary period of conscious work which always precedes all fruitful unconscious labor. Permit me a rough comparison. Figure the future elements of our combinations as something like the hooked atoms of Epicurus. During the complete response of the mind, these atoms are motionless, they are, so to speak, hooked to the wall; so this complete rest may be indefinitely prolonged without the atoms meeting, and consequently without any combination between them.

"On the other hand, during a period of apparent rest and unconscious work, certain of them are detached from the wall and put in motion. They flash in every direction through the space (I was about to say the room) where they are enclosed, as would, for example, a swarm of gnats or, if you prefer a more learned comparison, like the molecules of gas in the kinematic theory of gases. Then their mutual impacts may produce new combinations.

"What is the role of the preliminary conscious work? It is evidently to mobilize certain of these atoms, to unhook them from the wall and put them in swing. We think we have done no good, because we have moved these elements a thousand different ways in seeking to assemble them, and have found no satisfactory aggregate. But, after this shaking up imposed upon them by our will, these atoms do not return to their primitive rest. They freely continue their dance.

"Now, our will did not choose them at random; it pursued a perfectly determined aim. The mobilized atoms are therefore not any atoms whatsoever; they are those from which we might reasonably expect the desired solution. Then the mobilized atoms undergo impacts which make them enter into combinations among themselves or with other atoms at rest which they struck against in their course. Again I beg pardon, my comparison is very rough, but I scarcely know how otherwise to make my thought understood.

"However it may be, the only combinations that have a chance of forming are those where at least one of the elements is one of those atoms freely chosen by our will. Now, it is evidently among these that is found what I call the *good combination*. Perhaps this is a way of lessening the paradoxical in the original hypothesis.

"Another observation. It never happens that the unconscious work gives us the result of a somewhat long calculation *all made*, where we have only to apply fixed rules. We might think the wholly automatic subliminal self particularly apt for this sort of work, which is in a way exclusively mechanical. It seems that thinking in the evening upon the factors of a multiplication we might hope to find the product ready made upon our awakening, or again that an algebraic calculation, for example a verification, would be made unconsciously. Nothing of the sort, as observation proves. All one may hope from these inspirations, fruits of unconscious work, is a point of departure for such calculations. As for the calculations themselves, they must be made in the second period of conscious work, that which follows the inspiration, that in which one verifies the results of this inspiration and deduces their consequences. The rules of these calculations are strict and complicated. They require discipline, attention, will, and therefore consciousness. In the subliminal self, on the contrary, reigns what I should call liberty, if we might give this name to the simple absence of discipline and to the disorder born of chance. Only, this disorder itself permits unexpected combinations." (1913, pp. 392-394.)

Enough for the historical dating and exposition of the trial-and-error theory of creative thinking.* Whatever dominance it may have had in 1910 has been covered up in recent decades under the Gestalt treatment, which was explicitly hostile to the Thorndike (1898) description of problem-solving in terms of overt trial and error, although not explicitly contradicting a notion of a covert mental trial and error of the kind described by Bain, Mach and Poincaré.

* Souriau (1881) was another early exponent of the role of chance in thought and invention. See also Thurstone (1924). Emphases upon the importance of the fortuitous in experimental science are very numerous, as shown in the citations provided by Barber (1952), Aubert (1959) and Campbell (1959). Recent emphases upon the role of trial-and-error in logical and mathematical deduction occur in Quine (1947, 5-6) and Polya (1945, 1954).

(Gestalt psychologists were, of course, hostile to the associationist psychology with which both Bain and Mach were allied.) There is no essential contradiction of psychological fact here. As Woodworth and Schlosberg (1954, p. 823) point out in discussing Wertheimer on thinking, "it is trial and error in the nonderogatory sense, however, if any leads which suggested themselves and were tried out (or thought through) proved to be blind alleys." The protocols of human and animal problem-solving provided by the Gestaltists are easily reconciled to the model, provide ample evidence of both fortuitous solutions and misleading insights, as the studies of set in problem-solving illustrate. Thus while the pervasive influence of Gestalt psychology has led to the temporary eclipse of the blind-variation-and-selective-survival model, there is judged to be no tangible disagreement to be resolved.

Another prevalent orientation antithetical in spirit to the trial-and-error model can be called the "mystique of the creative genius and creative act." This is a deeply rooted tendency, related to our bias toward causal perception (e.g., Heider, 1944; Michotte, 1946), to see marvelous achievements rooted in equally marvelous antecedents. It takes the form of the "fallacy of accident" and of "*post hoc ergo propter hoc*." Let us suppose that a dozen equally brilliant men each propose differing guesses about the unknown in an area of total ignorance, and the guess of one man proves to be correct. From the blind-variation-and-selective-survival model this matching of guess and environment would provide us with new knowledge about the environment but would tell us nothing about the greater genius of the one man—he just happened to be standing where lightning struck. In such a case, however, we would be tempted to look for a subtle and special talent on the part of this lucky man. But, for the genuinely unanticipatable creative act, our "awe" and "wonder" should be directed outward, at the external world thus revealed, rather than directed toward the antecedents of the discovery. Just as we do not impute special "foresight" to a successful mutant allele over an unsuccessful one, so in many cases of discovery, we should *not* expect marvelous consequents to have had equally marvelous antecedents. Similarly, in comparing the problem-solving efforts of any one person; from the selective survival model it will be futile, in the instance of a genuinely innovative achievement, to look for special antecedent conditions not

obtaining for blind-alley efforts: just in so far as there has been a genuine gain in knowledge, the difference between a hit and a miss lies in the selective conditions thus newly encountered, not in talent differences in the generation of the trials.

This is not to deny individual differences in creative intellect. Indeed, the blind-variation-and-selective-survival model of creative thought predicts such talent differences along all of the parameters of the process. This is to emphasize, however, that explanations in terms of special antecedents will very often be irrelevant, and that the causal-interpretative biases of our minds make us prone to such over-interpretation, to a *post-hoc-ergo-propter-hoc* interpretation, deifying the creative genius to whom we impute a capacity for direct insight instead of mental flounderings and blind-alley entrances of the kind we are aware typify our own thought processes. Ernst Mach notes our nostalgia for the directly-knowing genius: "To our humiliation we learn that even the greatest men are born more for life than for science in the extent to which even they are indebted to accident." (1896, p. 175.)

What are the ways in which thinkers might be expected to differ, according to the trial-and-error-model? First, they may differ in the accuracy and detail of their representations of the external world, of possible locomotions in it or manipulations of its elements, and of the selective criteria. Differences in this accuracy of representation correspond to differences in degree of information and intelligence. Second, thinkers can differ in the number and range of variations in thought-trials produced. The more numerous and the more varied such trials, the greater the chance of success. Bain has emphasized the role of fanaticism or extreme dedication in producing large volumes of such explorations. Bain, Mach and Poincaré have emphasized the role of advance preparation in assembling the elements whose blind permutation and combination made possible a wide range of trials. Many observers have emphasized the role of set and familiarity in reducing the range of variations, and have recommended ways of reducing trial-to-trial stereotypy, as by abandoning the problem for a while, going on to other things. Devices abound which are designed to increase the likelihood that all permutations be considered and are used by most of us, as in going through the alphabet in finding rhymes or puzzle words. There are no doubt age differences in the rapidity and uninhibited

range of thought-trial production. The sociology of knowledge makes an important contribution here: persons who have been uprooted from the traditional culturally given, or who have been thoroughly exposed to two or more cultures, seem to have the advantage in the range of hypotheses they are apt to consider, and through this means, in the frequency of creative innovation. Thorstein Veblen (1919) has espoused such a theory in his essay on the intellectual pre-eminence of the Jews, as has Robert Park (1928) in writing of the role of "the marginal man" in cultural innovation. (See also Seeman, 1956.) And more generally, it is the principle of variation which leads us to expect among innovators those of personal eccentricity and bizarre behavior. We can also see in this principle the value of those laboratories whose social atmospheres allow wide ranging exploration with great tolerance for blind-alley entrances.

The value of wide ranging variation in thought-trials is of course vitiated if there is not the precise application of a selective criterion which weeds out the overwhelming bulk of inadequate trials. This editing talent undoubtedly differs widely from person to person, as Poincaré (1908, 1913) has emphasized. With regard to selection-criteria, one further point should be made. Much of creative thought is opportunistic in the sense of having a wide number of selective criteria available at all times, against which the thought-trials are judged. The more creative thinker may be able to keep in mind more such criteria, and therefore increase his likelihood of achieving a serendipitous (Merton, 1949; Cannon, 1945) advance on a problem tangential to his initial main line of endeavor. A final area of individual differences in competence is in the retention, cumulation, and transmission of the encountered solutions.

It need not be expected that these dimensions of talent all go together. In organic evolution, the variation process of mutation and the preservation of gains through genetic rigidity are at odds, with an increase in either being at the expense of the other, and with some degree of compromise being optimum. Just so we might expect that a very pure measure of innovative range in thought and a very pure measure of rote memory might be even negatively correlated, as Saugstad (1952) seems to have found, and similarly for innovative range and selective precision. Such considerations suggest complementary combinations of talent in creative teams,

although the uninhibited idea-man and the compulsive edit-and-record type are notoriously incompatible office-mates.

Note regarding the individual differences thus described that while they do make creative innovation much more likely on the part of some individuals than others, they do not place the joys of creative innovation beyond the reach of any one. Indeed, looking at large populations of thinkers, the principles make it likely that many important contributions will come from the relatively untalented and undiligent, even though on an average contribution per capita basis, they will contribute much less.

A final type of objection to the blind-variation-and-selective-survival model of thought needs to be considered. This objection is to the effect that the domain of possible thought-trials is so large that the solution of a given problem would take an impossibly long time were a search of all possibilities to be involved, either through a systematic scanning of all possibilities where these are enumerable, or through a still more tedious random sampling of the universe of possibilities. Time and trial estimates thus based can be overwhelming, as Kurt Lasswitz's story *The Universal Library* (1901) dramatically illustrates. Other parodies of our model occur in literature as far back as Swift's portrait of the Academy of Lagado in *Gulliver's Travels* (1726, pp. 166-169). (Ley (1958) traces such ideas back to Lully ca. 1200.) Newell *et al.* (1958a, 1958b) refer in this vein to what they call the "British Museum Algorithm," i.e. the possibility of a group of trained chimpanzees typing at random producing by chance in the course of a million years all of the books in the British Museum. Such parodies seem effectively to reject the blind-variation-and-selective-survival model through a *reductio ad absurdum*. Needless to say, such a rejection is not accepted in the present paper. As a matter of fact, it is judged to be in the same class as parallel objections to the theory of natural selection in evolution. Similar features in these two instances make the accidentalist interpretation more acceptable.

(1) Neither in organic evolution nor in thought are all problems solved, nor all possible excellent solutions achieved. There is no guarantee of omniscience. The knowledge we do encounter is achieved against terrific odds.

(2) The tremendous number of nonproductive thought-trials on the part of the total intellectual community must not be

underestimated . . . think of what a small proportion of thought gets uttered, what a still smaller fragment gets published and what a small proportion of what is published is used by the next intellectual generation. There is a tremendous wastefulness, slowness, and rarity of achievement.

(3) In evolution and in thought, the number of variations explored is greatly reduced by having *selective-survival criteria imposed at every step*. Thus mutant variations on nonadaptive variations of the previous generation are never tested—even though many wonderful combinations may be missed therefore. Current developments of “heuristics” in logic and chess playing machines (Newell *et al.*, 1958*a*, 1958*b*) have a similar effect of evaluating all next-possible moves in terms of immediate criteria, and then of exploring further variations upon only those passing the screening of each prior stage. It is this strategy of cumulating selected outcomes from a blind variation, and then exploring further blind variations only for this highly select stem, that, as R. A. Fischer has pointed out (1954, p. 91) makes the improbable inevitable in organic evolution. This strategy is unavoidable for organic evolution, but can obviously be relaxed in thought processes and in machine problem-solving. However, the Pandora’s box of permutations opened up by such relaxation can be used to infer that, in general, thought-trials are selected or rejected within one or two removes of the established base from which they start. In constructing our “universal library” we stop work on any volume as soon as it is clear that it is gibberish.

(4) When we make estimates of the number of permutations which would have to be culled to obtain a given outcome, we often assume that problem-solving was undertaken with that one fixed goal in mind. This overlooks the opportunistic, serendipitous course of organic evolution and of much of creative thinking. The likelihood of a productive thought increases with the wider variety of reasons one has for judging a given outcome “interesting.” To neglect this opportunistic multiple-purposedness gives one a poor base for estimating the probability of encountering the one outcome hit upon and recorded. Thus when Newell *et al.*’s *Logic Theorist* (1958*a*, 1958*b*) sets out to prove the 60-odd theorems in a given chapter of *Principia Mathematica*, it may face a more formidable task than did Whitehead and Russell in generating them—if,

except for the dozen classic theorems reproduced, Whitehead and Russell were otherwise free to record every deduction they encountered which seemed "interesting" or "non-trivial." Wigglesworth (1955, p. 34) has noted this strategy on the part of "pure" scientists, in commenting on the relationship between pure and applied scientists in wartime: "In the pure science to which they were accustomed, if they were unable to solve problem *A* they could turn to problem *B*, and while studying this with perhaps small prospect of success they might suddenly come across a clue to the solution of problem *C*."

In presenting their case for adding "heuristics" to the program of the "Logic Theorist," Newell *et al.* have emphasized the inadequacy of blind trial-and-error. I do not, however, find any essential disagreement between their point of view and the one offered here. By adding "heuristics" they have made the mechanical thought processes more like those of human beings, both in adequacy and the type of errors made. They have obviated the protests of those such as Wisdom (1952) and Mays (1956) who, while conceding that machines could choose good moves at chess or solve logic problems, have found the machines failing to imitate life just in their orderly inspection of all possibilities. Newell *et al.* recognize that a machine which would develop its own heuristics would have to do so by a trial-and-error of heuristic principles, with no guarantee that any would work. They further recognize that possession of an effective heuristic represents already achieved general knowledge about the domain under search, and that adding to this general knowledge will be a blind search process. They might also agree that most heuristic devices will be limited to the specific domain of their discovery, and can only be extended to other domains on a trial basis. They would probably also agree that no problem-solving process will be "direct." The disagreements I have with their excellent paper on the processes of creative thinking (1958*b*) are thus minor matters of emphasis, but may be worth stating none-the-less, to further clarify the position here advocated.

They say, for example, "We have given enough estimates of the sizes of the spaces involved . . . to cast suspicion upon a theory of creativity which places its emphasis upon increase in trial-and-error." (1958*b*, p. 63.) My position would unequivocally state that *ceteris paribus*, the wider the range of trials and the greater the volume of

trials, the greater the chance of a productive innovation. Doubling the number of efforts very nearly doubles the chance of a hit, particularly when trials are a small part of the total domain. But they too recognize unconventionality as a necessary, if not a sufficient condition of creativity (1958*b*, p. 62). What they would validly stress, more than has been stressed here, is the very frequent tactical advantage of a trial-and-error of general strategic principles over a trial-and-error involving no classificatory effort nor attempt to use clues. The advantage of such a strategy depends upon the ecology, of course, but we are in general justified in expecting solutions to be nonrandomly distributed, and to show significant contingencies with prior clues.

Another minor point of disagreement may be mentioned. In their efforts to consider how a "Logic Theorist" might be programmed to learn a general heuristic from hindsight they propose that it keep a record of the outcomes of all past trials, successful and unsuccessful, in order to be able to scan its experience for general principles of strategy (1958*b*). Implementing this would put a tremendous strain upon memory storage, and would introduce a scanning process as time-consuming as the original search process which produced the record. The strategy of organic evolution is to keep a record only of what works, even at the expense of repeating its errors, and the general preponderance of wrong tries, plus memory glut and access problems, suggests a similar strategy for all knowledge processes. Heuristics can probably best be learned through a trial-and-error of heuristics, tried on new problem sets rather than old. But while I judge this a valuable insight from the evolutionary process, it represents a very minor criticism of their most admirable program of research.

SUMMARY

This paper has attempted to make the psychological and epistemological point that in all processes leading to expansions of knowledge, a blind-variation-and-selective-survival process is involved. Processes substituting for an overt trial-and-error are of course acknowledged, with vision and thought being treated in some detail. But each of these are interpreted as containing in their very workability wisdom about the environment obtained by the blind variation of mutation and natural selection. In addition, each contains a blind-variation-and-selective-survival process at its own level.

Supporting the effort to interpret all knowledge processes in this light has been an emphasis upon the tremendous gain in knowledge in the course of evolution and history, a gain which can only be explained by a continual break-out from the bounds of what was already known, a break-out for which blind variation provides the only mechanism available. There has also been an effort to root out a prevailing implicit belief in "direct" or "insightful" knowledge processes, a belief which the phenomenal directness of vision encourages.

REFERENCES

- ASHBY, W. R. (1952). *Design for a Brain*, Wiley, New York.
- AUBERT, V. (1959). Chance in social affairs, *Inquiry* (University of Oslo Press), 2, 1-24.
- BAIN, A. (1874). *The Senses and the Intellect* (1st ed. 1855) 3rd ed., Appleton, New York.
- BALDWIN, J. M. (1900). *Mental Development in the Child and the Race*, Macmillan, New York.
- BALDWIN, J. M. (1906). *Thought and Things or Genetic Logic, Vol. I: Functional Logic*, Macmillan, New York.
- BALDWIN, J. M. (1908). *Thought and Things or Genetic Logic, Vol. II: Experimental Logic*, Macmillan, New York.
- BALDWIN, J. M. (1911). *Thought and Things or Genetic Logic, Vol. III: Real Logic, Interest and Art*, Macmillan, New York.
- BARBER, B. (1952). *Science and the Social Order*, The Free Press, Glencoe, Ill.
- CAMPBELL, D. T. (1956a). Adaptive behavior from random response, *Behavioral Sci.* 1, 105-110.
- CAMPBELL, D. T. (1956b). Perception as substitute trial and error, *Psychol. Rev.* 63, 330-342.
- CAMPBELL, D. T. (1958a). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Sci.* 3, 14-25.
- CAMPBELL, D. T. (1958b). Systematic error on the part of human links in communication systems, *Information and Control* 1, 334-369.
- CAMPBELL, D. T. (1959). Methodological suggestions from a comparative psychology of knowledge processes, *Inquiry* (University of Oslo Press), 2, 152-182.
- CAMPBELL, D. T. and FISKE, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychol. Bull.* 56, 81-105.
- CANNON, W. B. (1945). *The Way of an Investigator*, Norton, New York.
- DITCHBURN, R. W. (1955). Eye-movement in relation to retinal action, *Optica Acta* 1, 171-176.
- DITCHBURN, R. W. and FENDER, D. H. (1955). The stabilized retinal image, *Optica Acta* 2, 128-133.
- DUNCAN, C. P. (1959). Research on human problem solving: 1946-1957, *Psychol. Bull.* 56, 397-429.
- FISHER, R. A. (1954). Retrospect of the criticism of the theory of natural selection in *Evolution as a Process*, (ed. by J. HUXLEY, A. C. HARDY and E. B. FORD), Allen & Unwin, London.

- GRIFFEN, D. R. (1958). *Listening in the Dark*, Yale Univ. Press, New Haven.
- HADAMARD, J. (1945). *The Psychology of Invention in the Mathematical Field*, Princeton Univ. Press, Princeton.
- HEBB, D. O. (1949). *The Organization of Behavior*, Wiley, New York.
- HEIDER, F. (1944). Social perception and phenomenal causality, *Psychol. Rev.* **51**, 358-374.
- JENNINGS, H. S. (1906). *The Behavior of the Lower Organisms*, Columbia Univ. Press, New York.
- KELLOGG, W. N. (1958). Echo-ranging in the porpoise, *Science* **128**, 982-988.
- LASSWITZ, K. The University Library (first published 1901), in *Fantasia Mathematica* (ed. by C. FADIMAN), Simon and Schuster, New York.
- LEY, W. (1958). Postscript to "The Universal Library," in *Fantasia Mathematica*, (ed. by C. FADIMAN), pp. 244-247, Simon and Schuster, New York.
- LORENZ, K. (1943). Die angeborenen Formen möglicher Erfahrung, *Z. Tierpsychol.* **5**, 235.
- LORENZ, K. (1951). The rule of Gestalt perception in animal and human behavior, in *Aspects of Form* (ed. by L. L. WHYTE), Pellegrin & Cudahy, New York.
- MACH, E. (1896). On the part played by accident in invention and discovery, *Monist* **6**, 161-175.
- MAYS, W. (1956). Cybernetic models and thought processes, *Proc. 1st Intern. Congr. on Cybernetics (Namur)*, Paris.
- MERTON, R. K. (1949). *Social Theory and Social Structure*, The Free Press, Glencoe.
- MICHOTTE, A. E. (1946). *La Perception de la Causalité*, Inst. sup. de Philosophie, Louvain, Etudes Psychol., vol. VI.
- NEWELL, A., SHAW, J. C. and SIMON, H. A. (1958a). Elements of a theory of human problem solving, *Psychol. Rev.* **65**, 151-166.
- NEWELL, A., SHAW, J. C. and SIMON, H. A. (1958b). The process of creative thinking, University of Colorado 1958 Symposium on Cognition, Rand Corporation report P-1320, 16 September 1958.
- PARK, R. E. (1928). Human migration and the marginal man, *Amer. J. Sociol.* **33**, 881-893.
- PIAGET, J. (1950). *Introduction à l'épistémologie génétique*, I: *La pensée mathématique*, II: *La pensée physique*, and III: *La pensée biologique, la pensée psychologique, et la pensée sociologique*, Presses Universitaires de France, Paris.
- PLATT, J. (1956). Amplification aspects of biological response and mental activity, *Am. Scientist* **44**, 181-197.
- PLATT, J. (1958). Functional geometry and the determination of pattern in mosaic receptors, in *Symposium on Information Theory in Biology* (ed. by P. P. YOCKEY, R. L. PLATZMAN and H. QUASTLER), Pergamon Press, pp. 371-398.
- POINCARÉ, H. (1908). L'invention mathématique, *Bull. Inst. gen. Psychol.* **8**, 175-187.
- POINCARÉ, H. (1913). Mathematical creation, in *The Foundations of Science* (ed. by H. POINCARÉ), Science Press, New York.
- POLYA, G. (1945). *How to Solve It*, Princeton Univ. Press, Princeton.
- POLYA, G. (1954). *Mathematics and Plausible Reasoning*, Vol. I: *Induction and analogy in mathematics*; Vol. II: *Patterns of plausible inference*, Princeton Univ. Press, Princeton.

- PRINGLE, J. W. S. (1951). On the parallel between learning and evolution, *Behaviour* 3, 175-215.
- PUMPHREY, R. J. (1950). Hearing, in *Symposia of the Society for Experimental Biology*, Vol. IV: *Physiological mechanisms in animal behavior*, Academic Press, New York.
- QUINE, W. VAN O. (1947). *Mathematical Logic*, Harvard Univ. Press, Cambridge.
- RIGGS, L. A., RATLIFF, F., CORNSWEET, J. C. and CORNSWEET, T. N. (1953). The disappearance of steadily fixated visual test object, *J. Opt. Soc. Amer.* 43, 495-501.
- RIGGS, L. A., ARMINGTON, J. C. and RATLIFF, F. (1954). Motions of the retinal image during fixation, *J. Opt. Soc. Amer.* 44, 315-321.
- SAUGSTAD, P. (1952). Incidental memory and problem-solving, *Psychol. Rev.* 59, 221-226.
- SEEMAN, M. (1956). Intellectual perspective and adjustment to minority status, *Social Problems* 3, 142-153.
- SIMON, H. A. (1957). *Models of Man*, Wiley, New York.
- SOURIAU, P. (1881). *Théorie de l'Invention*, Hachette, Paris.
- SWIFT, J. (1941). *Gulliver's Travels* (1726), Blackwell, Oxford.
- THORNDIKE, E. L. (1898). Animal intelligence: an experimental study of the associative processes in animals, *Psychol. Rev. Monogr. Suppl.* 2, No. 8.
- THURSTONE, L. L. (1924). *The Nature of Intelligence*, Harcourt Brace, New York.
- VEBLEN, T. (1919). The intellectual pre-eminence of Jews in modern Europe, *Politl. Sci. Quart.* 34, 33-42.
- WIGGLESWORTH, V. B. (1955). The contribution of pure science to applied biology, *Annals Appl. Biol.* 42, 34-44.
- WISDOM, J. O. (1952). Symposium: Mentality in Machines, in *Aristotelian Society Suppl.*, Vol. 26, *Men and Machines*, pp. 1-26, Harrison & Sons, London.
- WOODWORTH, R. S. and SCHLOSBERG, H. (1954). *Experimental Psychology* (rev. ed.), Holt, New York.

DISCUSSION

MINSKY: I can't accept this feedback as being positive. I am not sure whether it is negative or not. I would really like to know if you believe you are saying something constructive rather than anarchistic? Namely, it seems to me what you have said particularly in reference to Newell *et al.*, that you don't believe things are as bad as they make out in the British Museum algorithm. That is, the space isn't really so large, there aren't so many bad trials and generally speaking things are pretty good. However, if they are not, you have to use heuristics as they do. What is the constructive purpose in emphasizing the role of trial-and-error which is what we are trying to get rid of?

When a great man solves problems consistently—Now, you take ten Einsteins, and put them in a room, and have them consider the problem of a generalized field theory. One of them will get it and you say, "Well, he was lucky." This was the lucky Einstein. Surely you can't be serious. You know then that you will say, "Now, I will watch this Einstein a little more carefully." I submit to you that the bias people have for making casual judgments of *post hoc ergo propter hoc* is one of the great discoveries we have made and we put in all of our machines and we don't try and err any more about that.

CAMPBELL: The last point, I heartily agree with. Causality is totally unjustified from a philosophical point of view but it is one of our great animal inductions

found in the white rat and in college sophomores and elsewhere. If I emphasize that in this instance it is leading us into error, I still recognize it as a tremendous evolutionary achievement.

Now as to the more general criticism, I guess I would like to have Newell discuss this. It seems to me that what you are doing with heuristics is analogous to what has been done by vision and by thinking as I have described them: you are discovering empirically that a certain mapping is useful, and within that mapping you conduct a blind matching, or a search, or a scanning, which then is substituted for the more costly lower-level trial-and-error. Now why I should want to call this abbreviation and substitution process still "blind" reflects, I suppose, some antimetaphysical bias within my own personality.

But I would disagree with some of the notions which were explicit in your talk this morning. You spoke hopefully of "intelligent" computer programs which would solve problems "directly." With these hopes I do disagree, and do predict that in so far as any problem-solving problem goes beyond what it already knows, it will do so by a "stupid" scanning, matching, or exploration process. Yet I found nothing I could disagree with in Newell's presentation. He did not claim that blind search could be obviated other than by an abbreviated symbolic blind search, usable only because of an already discovered wisdom about the domain in question.

Now as for the problem of Einstein, the problem of genius. There are of course many reasons why some people are regularly creative and others not which fit in perfectly with the blind-variation-and-selective-survival model. But over and above the individual differences in degree of genius, there is still a great deal of chance. If you look at the new discoveries within your own field, within your own generation, I suspect that you will see people whom you know are dumb making marvelous discoveries. Perhaps if you haven't experienced this it is because you have been one of the lucky ones. (Laughter.)

MINSKY: I agree that is what we are saying. The question is, what you are saying—we are trying to minimize the search. What are you proposing as an alternative?

CAMPBELL: I do want to avoid claiming any productive contribution to this conference. This blind trial-and-error theme is my way of making use of what I have learned from you folks on previous occasions. I have heard nothing on this occasion which leads me to change my mind on this point even though it may be a trivial point.

Is Newell here so we can abbreviate the discussion this afternoon and get on to the other issues?

NEWELL: The impression I get is the following. You start with the problem and the first thing you try and do is to put structure in the problem. You will back off to the place where something is certain, namely simply how you could generate possible logic expressions, we could take that as given—at least you can put down some kind of letters. You put down a base model. You say at least this is given. You say given this you go from here and the solution is one of the things that is generated this way. Now all spaces for all real problems that one can generate are found to be very large. You find such numbers as 10^{120} . They have to be a very large space, one for which if this were all you knew, and for which if you then proceeded to vary from this point, for that problem as stated; it's easy to make the calculations you are not about to solve it. You could state a different problem such as that in chess to start from the present position to any mate position you will find it doesn't improve it very much. You will find it is so big you are not about to do it, but just knowing that all you can do

is generate new situations and ask whether this is a checkmate—look around for some more information. Given large lists of these, which are what we call heuristics, each of these cuts down the list a little bit. At some stage of the game unless this is a trivial problem, and a trivial problem means here that the mathematician has given us a neat algorithm for solving it, most of the time you find there is a state beyond which you have no more information from there on, and you must do nothing but vary. The interesting question, which is why I find myself not emphasizing the blind variation aspect of it—the interesting question is, how do you get from a big space to a little space? We don't know how humans do it and this is the thing we would like to find out. The content of our science consists of lists of heuristics telling how people get from big spaces to small spaces and the fact is one should not overestimate the fact that for very difficult problems we can't get the thing down very small and it takes a man a very long time to solve certain problems. One should not overestimate the amount of knowledge that is available to strip down big spaces into small spaces. Nevertheless, it is this difference from the big space to the small space, it seems to me, that constitute the science, and the thing we call intelligence. Beyond this it is trial-and-error, you don't have any information, so what the heck do you do? I don't know if this sheds any light on it. I don't see anything inconsistent with all of these. I would ask you to go back and take one of these other situations and begin making an estimate. This would be lots of fun. You know I am aware of the remark made a moment ago about no positive contribution, but seriously, if one really thinks some of these spaces aren't so big, it is in fact because some of us have been overlooking some things and I think it would be great sport to find out why the spaces are not quite as big as we think they are.

CAMPBELL: I think we all agree, looking at Ashby's machine and probably the Newell, Shaw and Simon machine before it had heuristics, that any effective problem-solving machine will have a lot more structure in it than did these machines, just because of the tremendous domain of possibilities. I guess many of us in the long run will agree that most of this structure will be introduced between Mark 1 and Mark 2 or Mark 50,000 and Mark 50,001. That is, it will be put in between models and thus will not be achieved by self-organization.

In effective problem-solving machines there is a tremendous amount of *given* structure. I know this is Roger Sperry's opinion of Ashby's quotation of Sperry. Sperry just doesn't recognize his own works at all, because he, in contrast to Ashby, emphasizes that the circuits in animals are not reversible: nerves for left arms will not work for right arms in most species, etc.

The point I would like to make is, in accepting as a brilliant contribution the emphasis upon heuristics and hierarchies of heuristics, is that moving from each level to each level higher and discovering new effective heuristics represents gained information about the domain (comparable to that which was already achieved by the invention of vision). In addition, this information is gained by a blind exploration process, and contains in its execution trial-and-error components. So that I guess the truism that I have advocated still remains a truism, although perhaps a trivial one for those who are trying to learn about particular effective short cuts.

JORDAN (*Systems Development Corporation, Santa Monica, California*): You mentioned one sentence there: of course there are individual differences—I have a feeling that by mentioning this sentence you sort of opened the back door and the whole problem came back, the problem you wish to eliminate. What is meant by individual differences?

If we mean by this systematic differences of performance, we either then have to postulate one of two things; one, let us categorize these individual differences in terms of people doing well, people being pedestrian and people being stupid in problem solving. Now you either have a common blind trial-and-error which all organisms have because that is the nature of organisms and then some people cope with it more effectively than others, so their performance is perceived under the measurement of individual differences, or you postulate another thing, every individual carries his own trial-and-error apparatus within him. Some are more efficient than others, so you have different kinds of trial-and-error.

So I might submit in both of these cases the problems of creativity and stupidity re-emerge and are solved because you have these. You can say there is no such thing as individual differences, this is a chance aggregate of hits. We are all making trial-and-error. Some of us hit and most of us do not. This guy was lucky and hit zero out of ten times.

I don't know how this would explain consistent performance, differentiation of performance, on intelligence tests which correlate pretty highly from 4 years onwards.

CAMPBELL: I do not want to deny individual differences and actually I have a section in the manuscript on it, which I did not have time to read. But the quotations read from Mach and from Bain emphasize individual differences in creative talent quite consistent with this model. The point I wanted to make was, however, our tendency to over-interpret achievement along this line. Actually Newell has commented on the same effect in his paper on creativity at the Colorado Symposium last spring. Our proneness is to over-interpret the antecedents of a lucky encounter.

Now as to the reliability of creativity, I suspect that in actual shops whose business is turning out inventions, reliability is considerably lower in terms of the number of patents awarded per year than it is for intelligence test performance. But this is a matter of evidence. Of course, from the trial-and-error model it follows that the man who can produce thought-trials faster, the computer with the greater speed, the computer with the greater memory, the computer that starts out already programed with the better trial-and-error heuristics will have a tremendous advantage.

CRITCHLOW (*J.B.M.*): In what way is random choice superior to program choice in selecting from many courses of action?

CAMPBELL: This is not a distinction which I argued. Both programed and random procedures could be blind in the sense which I wish to emphasize. However, let us consider a simple gadget, like Ashby's, which might have the alternatives of either systematically going through all possible states, or of randomly sampling from these. If the alternatives have an order in their storage, a systematic search may lead to more repetition in the sense of continuing to make very similar wrong responses and there might thus be cases in which random sampling of possibilities would be better; but they are both blind. The radar sweep is just as blind as it would be if you illuminated the phosphorescent pips at random. It is just as blind, even though it is systematic.

THE NATURAL HISTORY OF NETWORKS*

GORDON PASK

Systems Research Ltd., London, England

INTRODUCTORY REMARKS

IN THIS paper I shall put forward a pair of contentions, and examine some of their consequences. The first of these concerns any network which is given in nature or which appears as part of a "Black Box"⁽¹⁾ problem. The contention is that if an observer wishes to use any self-organizing potentialities the network may have, then he must look at the network as though he were a natural historian.

I am using the term "network" in a general sense, to imply any set of interconnected and measurably active physical entities. Naturally occurring networks, of interest because they have a self-organizing character, are, for example, a marsh, a colony of micro-organisms, a research team, and a man.

It is not so easy to say what I mean by a natural historian. Emphatically he is not a meticulous and classifying person. In choosing the name I had the interactive aspects of natural history in mind, the art of knowing about a rabbit run, almost by living the part of a rabbit, the skill of animal training—disciplined enough to permit its discussion—the search for similarities which are cogent within the network itself.

The idea of necessity also needs comment. We can, of course, look at a system in any way we choose, regardless of whether or not it is self-organizing. Thus, we can look at a man from the anatomical point of view and see a creature with two legs, bounded by its skin. Again, we might examine man like the sociologists and see a badly defined game player. The contention is that *in order to use the self-organizing character of a man we must become natural*

* The author wishes to acknowledge his association as Assistant Research Professor with Professor H. von Foerster (Electrical Engineering Research Laboratory) of the University of Illinois, during the writing of his paper. While at the University of Illinois the work was supported by the Information Systems Branch of the Office of Naval Research under Contract Nonr. 1834(21).

historians, which means, for the human system, that we must talk to it. In conversation the system appears to be bounded at one moment by the anatomist's skin, and at the next moment, by its region of influence upon other men in society. Typically a natural historian must change his viewpoint to suit a changeful system.

DIFFERENT METHODS OF OBSERVATION

In order to express these notions precisely we must examine more familiar ways of observing networks and compare these with the natural historian's strategy.

Any pattern of activity in a network, regarded as consistent by some observer, is a system. Certain groups of observers, who share a common body of knowledge, and subscribe to a particular discipline, like "physics" or "biology" (in terms of which they pose hypotheses about the network), will pick out substantially the same systems. On the other hand, observers belonging to different groups will not agree about the activity which is a system.

I shall call a body of knowledge, in which statements are related in a common language, a reference frame.⁽²⁾ For the observer who adopts it, a reference frame determines the kind of enquiry which is relevant (and thus, the set of physical attributes, of the activity in a network which it is pertinent to observe).

Ultimately, observations and experiments are conducted in order to control the activity in a network. Some observers, a category which will be defined as *specialized observers*, wish to control this activity by discovering the "truth" about how it occurs. They experiment, on the assumption that a sufficiently complicated "truth" is invariant, by trying to identify observable behavior with a hypothetical system. Now any real observer is limited by the number of states he can usefully distinguish in an experiment, and in general he will not be able to identify a sufficiently complicated hypothetical system in any direct fashion with the measurable activity. Rather, he will assimilate results from many experiments, each of which validates only a sub-system of the hypothetical system he assumes to exist. But assimilation is only possible if the experimental observations can be compared and transformed with a well-defined composition rule. It is no accident that the measurable attributes deemed relevant in different reference frames prove incomparable, which implies that results from experiments in

different reference frames are not assimilable. Because of this a *specialized observer* necessarily elects to experiment within a single reference frame. Thus, any specialized observer who examines a network will discern only those systems which are manifest:

- (1) By changes in a set of relevant variables, or equivalently
- (2) Which are composed of the unit elements or components which may be defined if the functional equations of the activity, appropriate for the particular reference frame, are reduced to canonical form.⁽³⁾

Systems may also be controlled by interacting with them, and this technique is used by a natural historian. In general, he assumes that the "truth" about the system is not invariant (otherwise he would have examined it like a specialized observer), and his experiments aim either to maximize future interaction or to achieve some more specialized objective (like making a system called an elephant, get up on its hind legs). The natural historian, since he is not seeking the absolute, adopts whatever relevance criteria allow him to achieve interaction.⁽⁴⁾ A specialized observer sees him skipping illogically from one frame to another; for example, at different stages in the training process he may "feed" and "entice" the elephant, which are procedures appropriate to strictly incomparable models of the system. As a result of his experiments the natural historian may be able to make assertions about how to interact—like "Give it a bun if its trunk is drooping," or "Pat the creature on the head each day"—but these must not be confused with truths about the system in the previous and rigorous sense. Giving a bun to and patting an elephant, both of which induce it to stand upright, are not procedures comparable with changing the pressure and the temperature of a gas, both of which make it change volume. The laws of gaseous behavior are expressed in a single reference frame. The laws of elephant behavior are not. Thus, a natural historian cannot say anything precise about the way that elephants (or other systems) work. He makes comments only about his interaction.

While admitting this limitation, I believe that a natural historian can answer all of the enquiries it is either legitimate or useful to make about a self-organizing system. The natural historian's language is appropriate for discussing behavioral characteristics some of which are vague, some of which (like the redundancies and stabilities described by McCulloch^(5,6) and the habituation

described by Ashby⁽⁷⁾), are firmly related to topological parameters of the network, and a few of which like "differentiation" and "memory" will be examined in this paper. The natural historian is unable to say where these behavioral characteristics reside in the network or how they are manifest, but questions of this kind are probably meaningless in the self-organizing context.

Mathematical Representation

It would be inappropriate, in this paper, to discuss the mathematical work which is being done by A. Mullin at the University of Illinois and which will elaborate these ideas. However, I shall present a few descriptive structures with the provision that a more elegant formulation will emerge when the optimum mathematical technique for dealing with the observer and network problem has been determined.

Specialized Observation

A model or hypothetical system $J_j = (U_j, G_j)$ is a set U_j^* of points U in a phase space $\mathcal{U}$ together with a finite group G_j of transformations $F \subset G_j$.

Clearly the models J_j with $(j = 1, 2, \dots)$ are consistent, due to their group character, and are elements of an hierarchy, the coherence of which depends upon the adoption, by a number of specialized observers, of a certain "Composition Law." In the present discussion we shall assume that this "Composition Law" is matrix multiplication and that the groups G_j have thus the usual connotation.

In this hierarchy the lowest hypothetical models will be determined by cyclic groups generated as the powers of a single transformation such as $G_i = F_i F_i^2 \dots F_i^{c-1}$ with $F_i^c = I$ the identity transformation and $(i = 1, 2, \dots)$.

The model $J_i = (U_i^*, G_i)$ is thus a model of a stable system, for any state $U \subset U_i^*$ will be repeated in c units of activity, in general, in c observational intervals.

In any models J an observer is able to describe, in a way which is unambiguous to other observers adopting his composition law, those features which are kept invariant by $F \subset G$. However, in order to say that this description is the "truth," the model to which it refers, and in the simplest case a model like J_i must be experimentally tested.

Before examining the experimental procedure, let us note that as a result of this testing there will be two disjoint subsets of hypothetical systems in the hierarchy, and thus two disjoint hierarchies; one being built up from experiments which yielded results confirming the existence of hypothetical systems, the description of which is "true," and the other including plausible but experimentally unconfirmed models.

An experiment is an attempt to make an identification Ω_i between the points U in a hypothetical system and the states of a real network. As a result of this an observer may also relate the abstract transformations F to real mechanisms in a network or alternatively to real stimulus procedures which he can use to modify the state of the network. If identification is possible for any J_i , the system J_i is said to exist in the network.

Because of his limitations an observer is unable to make any experiment he pleases. The restrictions are of two kinds. First of all, if we regard Ω_i as a mapping between a vector of observed values of measurable attributes of a network and a state of the hypothetical model, the mapping is not isomorphic, but is many to one. However, in order to satisfy the requirements of identification it must preserve the Composition Law of the observer. Thus, Ω_i is a homomorphism. Secondly, identification implies a qualitative decision to regard only certain attributes of a network as relevant. These chosen relevant attributes determine the observer's reference frame. Thus, in the reference frame α a set of real attributes, say $x_1, x_2, \dots, x_n$ are examined, so that observations of states of networks are vectors $X_t, X_{t+1}, \dots$ at instants $t, t+1, \dots$, and the real mechanisms are transformations $A_1, A_2, \dots$. Similarly, in the reference frame β a set of attributes $y_1, y_2, \dots, y_m$ is regarded as relevant and observations are the vectors $Y_t, Y_{t+1}, \dots$ and real mechanisms are $B_1, B_2, \dots$.

If α and β are distinct it is clear that the set of all experiments performed in α which affirmed the existence of a hypothetical system will determine a sub-hierarchy α^* included within the previously defined "true" hierarchy. We define reference frames α and β so that the variables x_ξ and y_χ are incomparable for $\xi = 1, 2, \dots, n$ and for $\chi = 1, 2, \dots, m$, in the domain of the observer's composition law. Because of this α^* and β^* are disjoint.

Two points must be mentioned. First, it is likely that observers

have selected relevant attributes, merely because affirmative results derived from measuring relevant vectors were self consistent, being in a sub-hierarchy α^* , but were discovered inconsistent with, or unrelated to, those derived from measuring different relevant vectors which would be in a disjoint sub-hierarchy β^* . Secondly, it is clear that reference frames $\alpha, \beta \dots$ sub-hierarchies $\alpha^* \beta^* \dots$ and even the "true" sub-hierarchy are defined with reference to knowledge at the moment. The concepts are useful, since we are limited observers. However, it is always possible that reference frames will be rendered indistinct, that we shall adopt a new composition law which relates previously incomparable quantities and that the region of "truth" will extend. In particular, the distinction which will be made in a moment between specialized observers and natural historians would disappear if observers were able to know everything about a network.

We can represent the simplest experiments of a specialized observer as shown in Fig. 1. In the particular structure of Fig. 1 a

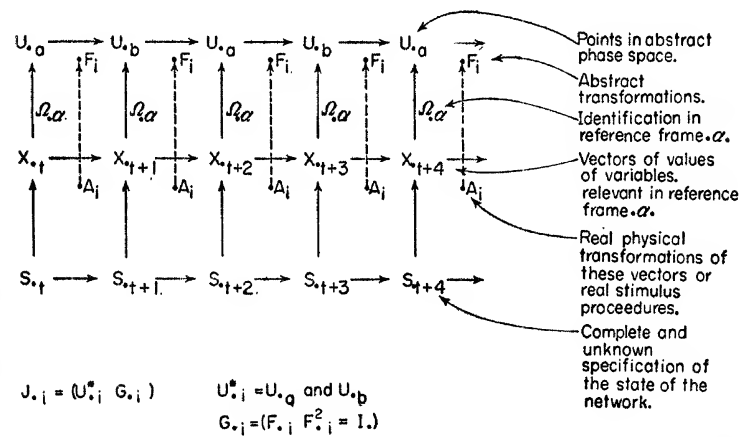


FIG. 1.

hypothetical model is identified in a reference frame α since the abstract state U_α is identified with X_i the abstract state U_b with the X_{i+1} and F_i with the mechanism A_i .

The actual states of the network, namely, $S_t, S_{t+1}, \dots$ and so on are accessible only to an observer able to know everything, and that is, for the present discussion at least, “unreal.” The mapping from

reference frame, U_i is not equal to $U_{i+\tau}$ for any finite τ , any identification, and any model J . Ashby⁽¹⁰⁾ calls them systems where the "truth" is changeful. It is no longer meaningful to make enquiries, as we tacitly do when adopting a reference frame, on the assumption that some descriptive "truth" remains invariant.

Natural Historians

Non-stationary activity in a network may be quite tractable for a natural historian because he is at liberty to relate entities which are incomparable to a specialized observer and to run back and forth through the dividing planes of the illustration. So far as the natural historian has a reference frame, it is simply the context of his *own interaction* with the network. Unlike the specialized observer, the natural historian has few preconceptions about a composition rule or about what entities or situations are equivalent.

Indeed, in the simplest case, when the natural historian is merely trying to "interact" and "make conversation"⁽¹¹⁾ with some system in the network, his strategy is to discover a set of composition rules and equivalence relations such that if he assumes them interaction will be favored.

We thus suppose the existence of composition rules $E_1, E_2, \dots$ which the natural historian is able to distinguish and to understand, but is not necessarily able to describe and equivalence relations $R_1, R_2, \dots$ of the same calibre. Initially, he chooses some E_i according to his view about the character of the network as a conversation partner—not according to his knowledge of its structure; and he also chooses some R_i , such that two R_i related situations have the same significance with reference to his interaction. He now seeks to modify these assumptions, as a result of interaction, so that interaction is favored.

Let V_μ, V_η be the names given to any state of the network which the natural historian is able to recognize. Let $P_i, P_n, \dots$ be the names given to any procedures which the natural historian can use to modify the activity in the network. According to the analogy of a conversation $P_i(R_i)P_n$ are a pair of equivalent gambits and $V_\mu(R_i)V_\eta$ are a pair of replies with the same meaning, so far as the second partner is concerned. The process of searching, which gives rise to a sequence $R_1 \rightarrow R_2 \dots \rightarrow R_p$ and (because alterations in R induce alterations in E) a sequence $E_1 \rightarrow E_2 \dots \rightarrow E_p$ is intuitively

familiar. The object of the search is to discover R^* and E^* such that

$$V_\mu(E^*)P^c(R^*)V_\mu$$

where

$$V_\mu(R^*)V_\eta \text{ if } (V_\mu(E^*)P_l)R^*(V_\eta(E^*)P_n)$$

when

$$P_l(R^*)P_n$$

Call the above consistency or predictability condition “&” and note that many pairs E^*R^* will permit satisfaction of “&” for different V_μ and P_l . Let b_γ equal the number of names V_μ and P_l for which a particular pair $R_\gamma^*E_\gamma^*$ permit satisfaction of “&.” In general the pair

$$E_\delta^*R_\delta^*$$

is preferred to the pair

$$E_\gamma^*R_\gamma^*$$

only if $b_\delta > b_\gamma$.

The searching sequence, which represents the interaction, will ideally approach an $E_\varepsilon^*R_\varepsilon^*$ such that b_ε is greater than any of the previously obtained values.

If the search process is one-sided so that the natural historian changes his viewpoint a great deal but has little effect upon the network, convergence toward a high valued b_ε is unreliable and inefficient. We shall later examine conditions (which always apply if the network is self-organizing rather than merely intractable) in which (because the network is modified by a “reward” under the natural historian’s control) an appropriate “rewarding strategy” will achieve rapid and efficient convergence.

Second Contention

The second contention refers to networks which are not given in nature, but which are deliberately built, so as to foster any self-organizing systems which appear. Oddly enough, there are conceptual difficulties which force us to look even at these constructed models in the manner of the natural historian. The contention is that these difficulties are not apparent in the abstract formulation but appear when it is embodied in any physical model such as a network.

It will be convenient to discuss the issue by formulating such an abstract model and then constructing one of the many physical realizations.

AN ABSTRACT MODEL OF A SELF-ORGANIZING SYSTEM

We require a number of concepts for building an abstract model.

(1) A space of arbitrary dimension in which a network is defined by asserting connectivity between pairs of points. Let us envisage this space filled with an initially homogeneous but malleable material M .

(2) A currency, which may be identified with energy, which is conserved on the average. The conservation conditions make measurement possible and will be secured if we have a definite rate θ at which currency or energy flows through the space.

(3) A set of currency seeking servomechanisms. We can identify these elements with von Foerster's Maxwell Demon servomechanisms,⁽¹²⁾ or equally with the catalysts which Prigogine and others^(13,14) describe as inducing open reaction systems in a stationary state network. They are, thus, non-linear amplifiers or oscillators with a local energy or currency store. Let us assume these servomechanisms exist at uniformly distributed points in the space. The only sense in which any one of these servomechanisms can increase the currency it has available is by influencing the activity of the others. This it may do by transmitting a trial or signal, which other elements sense, and the servomechanism in question is informed of the state of the network by receiving the effect exerted by the trials of other elements at its own input location. However, in making a trial or sending a signal, each servomechanism loses currency—in other words—there is a definite cost per trial.

(4) A set of rules determining the change in signal and currency connectivity induced by activity in the space. These are conveniently expressed by the signal impedance characteristics of the malleable material M .

(a) Consider a pair of points ij in M and a path m_{ij} connecting ij in M . Suppose a signal traverses the path m_{ij} at t , the signal impedance of m_{ij} at $t + 1$, say $\rho_{(ij) t+1}$ is less than the previous signal impedance $\rho_{(ij) t}$, the decrease being due to passage of a

signal at t . In the absence of further signals along this pathway, the signal impedance will increase and reach its original value in some finite interval τ , then $\rho_{(ij) t} > \rho_{(ij) t+1}$ only if a signal passed along m_{ij} in some preceding interval less than τ . A pathway may thus be defined simply as an m_{ij} connective in M where the signal impedance is greater than the average impedance in the ij neighborhood. Clearly, pathways, structures of signal connectivity, or networks arise necessarily if the servomechanisms are active and continue to exist only if they are used.

(b) When a servomechanism is active it uses currency. Let some servomechanism at a point i be active so that a more than average flow of currency must occur in M along pathway m_{ij} . When such a flow of currency occurs the currency impedance of m_{ij} , say $\kappa_{(ij)}$ will increase. If the flow of currency is greatly decreased, due to a suppression in the activity of the servomechanism at i , the currency impedance $\kappa_{(ij)}$ decays over a finite interval.

It may be argued:

- (i) that activity must occur, for the currency available must be used.
- (ii) that this activity will give rise to some kind of connectivity in M .
- (iii) that an active set of connections imply local activity.

Since local activity engenders local currency depletion, a new currency distribution and thus a new activity distribution is induced by the existence of the original connectivity.

It is clear that uniform connectivity in M together with uniform activity of the servomechanism elements is, in general, an unstable equilibrium, because in these conditions the system is searching for and is maximally sensitive to, any disturbance which will interrupt the uniformity. However, such a uniform state, which I shall call a λ state, may be shown to be the only stable state if the system is closed (except with reference to currency), so that no disturbance can occur. We may obtain this result by applying von Foerster's⁽¹⁵⁾ Multiservo Convergence Theorem to the *servomechanism elements*, when, assuming uniform connectivity, the representative points of the *ensemble of servomechanisms* must converge to a fixed point. But, if this tendency should occur, then whatever the initial connectivity of a closed system of this kind, the terminal connectivity will approach a uniform pattern. In such a variable connectivity

system there are indeed many different connectivities equivalent in that they are all uniform; thus the system will approach one of a set of λ states. The limit case of a system, such as Dr. McKay's trial making servomechanism,⁽¹⁶⁾ which has fixed connectivity and approaches a unique λ state is, however, included by the formulation.

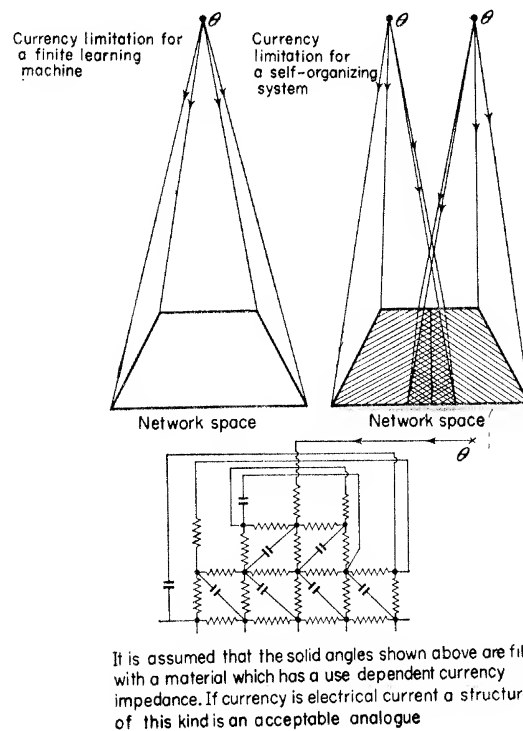


FIG. 3.

Clearly, if the state of the system is coupled to parameters of an environment and the state of the environment is made to modify parameters of the system, a learning process will occur. Such an arrangement will be called a *Finite Learning Machine*, since it has a definite capacity. It is, of course, an active learning mechanism which trades with its surroundings. Indeed it is the limit case of a

self-organizing system which will appear in the network if the currency supply is generalized.

Suppose that the network space is indefinitely extensive and that instead of limiting currency flow to θ per network space we restrict the flow to θ per unit volume of the network space. In this case there is an advantage to be gained in terms of the competition for currency between the servomechanism elements, if these elements co-operate. In other words, a set of servomechanisms is at an advantage if its activity extends the connected region in the network space. The extension will only be limited by the gain of the servomechanisms and the currency available. In realizable systems an active connected region moves around the network space capturing uninvolved servomechanisms. Such a system will be called *an abstract self-organizing system*.⁽¹⁷⁾ Since we cannot satisfactorily demark the active system, the inactive region in the network, and the environment; closure cannot be applied. However, if it could, there would be an indefinitely large number of λ states and these are approached as closure is approximated.

When related to a specific environment, this is a learning machine but not a finite learning machine, since the extent of the active system depends upon the external conditions. The relation of such a self-organizing system to the finite learning machine is indicated in Fig. 3.

Physical Construction of the Model

When any physical model is constructed, its maker has to accept certain essential constraints inherent in the medium. The existence of non-linearity in any real amplifier, the thermal coefficients of any real resistor are, for example, essential constraints. However, in building most models it is possible to select only one set of restrictions as being relevant to the action of the physical artefacts. Thus, when an electrical analogue computer is used to embody some abstract mechanism we say that an electrical model has been constructed, meaning that in realizing this abstraction we take account of the electrical model, but that we disregard, as not being relevant, the mechanical and thermal constraints inherent in the computer. Indeed, the computer is designed with this object in view, and a different computer might have been designed to embody abstractions in a mechanical model, electrical effects being discounted.

In terms of our convention, a computer analogue is a physical model so designed that its activity is explicable in a single reference frame, in the case I have cited, an electrical reference frame. In terms of engineering, such a mechanism is designed with components, like valves and resistances, which have a well specified function. A valve, for example, accepts only a electrical input and provides an amplified electrical output. If it also responds to temperature or vibration, it is to this extent a bad valve. The logical simplicity of the computer model is a consequence of being able to put one's finger upon a component which performs a known function and to reject the imperfections as irrelevant.

When trying to construct the physical model of an abstract self-organizing system we are beset with a peculiar difficulty. Not only are there many possible mechanisms which embody the abstract concept, but any mechanism we choose will embody it in an ambiguous manner. The logical requirements force us to use media such that, when a physical model is constructed, we cannot specify components which have a well defined function, and we cannot separate inputs and outputs into a set which are relevant and a set which may be discounted.

It is inherent in the logical character of the abstract self-organizing system that all available methods of organization are used, and that it cannot be realized in a single reference frame. Thus, any of the tricks which the physical model can perform, such as learning and remembering, may be performed by one or all of a variety of mechanisms, chemical or electrical or mechanical.

Thus, however much we try, we cannot achieve an electrical model or a mechanical model or a chemical model of a self-organizing system. Any physical model necessarily includes them all in varying degrees, and to a specialized observer they will appear distinct and incomparable, although a natural historian will be able to see them as equivalent.

To emphasize this point let us consider the process of differentiation. Suppose that at an early stage in its development the system has learned the advantage of having "individuality" in the sense that it has developed a primitive mechanism using specific substances, for example—proteins—to tag each "individual entity." Later the system learns about the existing primitive mechanism and evolves a more efficient device whereby "individual entities" are

spatially separated regions connected, distinctively, with fibers as in a nervous system. The primitive and the efficient mechanisms are functionally equivalent to a natural historian who regards "individuality" as a "behavioral characteristic" but incomparable to a specialized observer for whom "individuality" is unmeasurable. (In the network I shall describe, it happens that regional connectivity is given, but it is possible to distinguish at least two equivalent mechanisms which are developed for mutual inhibition, one acting by energy depletion and one which involves a specific connectivity.)

These ideas can be placed on a firm theoretical foundation by considering the system as it approaches a λ state. In this limiting condition a specialized observer sees a meaningless activity from which he can only infer the existence of a chance machine. A natural historian, on the other hand, sees a system which is maximally sensitive to any disturbance and liable to develop any one of many equivalent structures according to the disturbance it happens to appreciate. However, I shall not pursue this theoretical argument. Having made the point that we must view constructed networks as though we were natural historians, just as we have to view the self-organizing networks given in nature, it will be more instructive to examine the behavior of a real mechanism.

A Particular Physical Model

I shall describe a model⁽¹⁸⁾ in which the "currency" of the abstract system becomes electrical energy and signals may be thought of, initially, as electrical impulses. It will be convenient to describe the physical representation of servomechanism elements and of malleable material separately.

(1) In the model a "Maxwell Demon" servomechanism is an energy dependent trial making amplifier. Due to a mechanism which involves a refractory interval, it may distinguish its own output from the output of other elements, or the delayed effect of its own output acting as an input.* The element produces electrical output impulses called trials, because in the first place sufficient electrical energy is dissipated to modify the state and thus the connectivity of the surrounding material, and secondly, because

* In this respect the model is similar to the scheme discussed in ref. 19.

the impulse, transmitted through any existing connections may affect the trial making activity of other elements.

Each element has an electrical reservoir in which trial making energy is accumulated. Occasionally, the stored energy is dissipated by autonomous trial. In general, however, the sequence of trials is modified by inputs received, at a much lower energy level, from other elements in the sense that an input stimulates the occurrence of a trial. The gain of the element, as an amplifier, is a function of the average difference between input and trial energy, and in practice, we may look upon any element as a servomechanism which is seeking to maximize interaction, subject both to energetic constraints and those imposed by the connectivity built up as a result of the previous activity.

(2) To clarify the presentation I have separated the energy supply network from the connectivity or signal network.

As shown in Fig. 4 the trial making amplifiers receive energy—in this case electrical current—from a resistance-capacitance network into which a current θ passes at one or more points of symmetry.

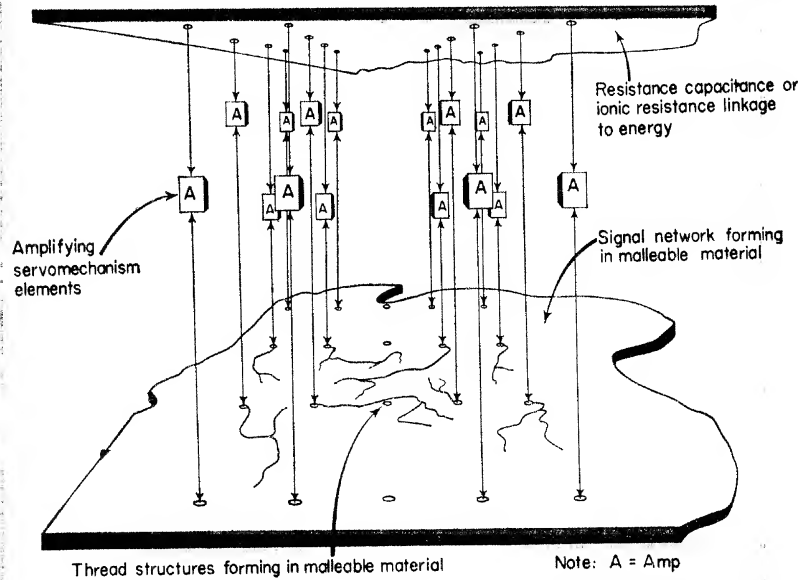


FIG. 4.

The resistance-capacitance network may be loaded as a result of excessive trial making activity on the part of a particular amplifier. If so, a local depletion will occur thus reducing the effective source-potential and the effective gain of the amplifier concerned.

The signal network is built up as a result of current passed by these amplifiers through a solution of ferrous sulphate, which is the malleable material *M*. As shown in Fig. 4, the current path is completed via a set of electrodes.

The electrode associated with each amplifier may act either as a source or a sink of d.c. current, according to whether a trial is or is not being made. If a trial is being made the amplifier also produces an a.c. signal at its electrode, which may be received by any electrode which is acting as a sink for d.c. current.

The solution itself is moderately conductive to the a.c. and the d.c. signals. However, if a d.c. current is passed between a source and a sink, a very low-resistance metallic thread develops from the sink along the line of maximum current, and gradually an entire network of threads is built up. The line of maximum current, where a particular thread develops, will depend upon the electrodes which are energized and also upon the existing network of threads since these, being of low resistance, act as extensions of the point electrodes. Thus, the network of threads not only distributes the a.c. signals which deliver inputs to the elements, but determines the further development of the network itself.

Once a thread is formed, there is a tendency for it to dissolve, due to a local acidity. A stable thread is thus in a dynamic equilibrium determined by the competition of a building up and a dissolving back process. A thread exists as a stable entity only if it is passing sufficient current to keep it intact and if the current is appropriately distributed. Clearly the distribution which *is* appropriate depends upon the entire network and its activity. In general, the network of threads determines the environmental parameters in which any particular thread develops and any particular thread determines the environmental parameters in which a small segment develops. Thus, the natural history of this network presents an over-all appearance akin to the natural history of a developing embryo or that of certain ecological systems.

Some of the mechanisms used in development are illustrated in Fig. 5.

(1) Shows a thread developing between electrodes d and e . It develops by a process of successive trial, nearly all the terminal trial threads being abortive.

(2) Shows the introduction of a further electrode f , as a result of which the thread may bifurcate.

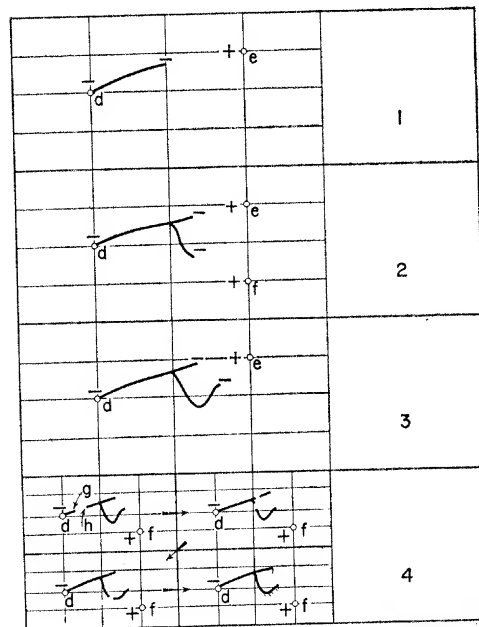


FIG. 5.

(3) Shows the development of the thread after bifurcation, but with only d and e energized. The effect of having previously energized f is apparent.

(4) Shows what will happen if, either due to instability or mechanical injury, the thread is split. The point g , being relatively negative, builds up new thread whilst the point h , being relatively positive, suffers dissolution. The process gives rise to regeneration of the thread as a whole which occurs up the branches de and df , even though only d and f are switched on. The existence of the thread has transformed the field which would have induced regenerative

development along df only into a field which induces development of de and df .

If a subset of the electrodes are associated with the output connections of sensory devices which in turn receive an input from an environment, and if a further subset of these electrodes are associated with devices able to effect the environment the network will interact and change state, seeking dynamic equilibrium with reference to the environment. The extent of the active region, produced in the network as a result of this search, will depend upon the informational variety of the environment and the value of θ (Fig. 6).

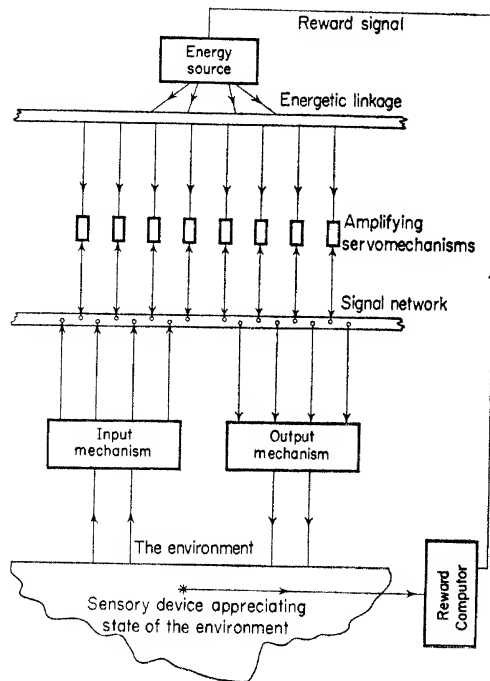


FIG. 6.

The adaptive process will lead to some system which can interact, in a stable fashion, with the environment. However, suppose that an observer tags a subset of the possible stable relationships as

desirable; in particular, suppose he wishes the system to act as a control mechanism, which achieves some state of affairs in the environment. He can train the system to adapt specifically in this way by controlling θ . Even a procedure such as increasing the value of θ only if the desired state is achieved will lead by natural selection to a structure which is a control mechanism aiming to achieve this state.

The over-all energy, or over-all currency, variable θ is here being used as a reward variable. This use of the term "reward" is, perhaps, unusual, and is certainly distinct from "rewards" which imply that a certain rewarded action becomes more probable. In the present case, when the network is rewarded, we mean that it is given permission to develop, that more of the constructional material may be used for making threads, and that more amplifiers may be included in the signal network and as part of the system. However, no restriction is placed upon the kind of development, which depends upon the existing structure.

A specialized observer would find this an unsatisfactory learning machine, because although it will learn what he wishes, he cannot tell how it learns, how to reward it, or how large it is. Before considering how a natural historian might administer a reward (according to the present contentions, with greater success), I should like to exhibit a few more characteristics of this model.

(a) We have already seen that given appropriate surroundings (namely a world of ferrous sulphate liberally bespattered with amplifiers) the system could extend wherever there is energy. Such a world is unlikely, so we enquire what will happen when a developing system reaches a boundary such that there is no more ferrous sulphate. This is the most primitive possible demarcation of an environment. Equally, of course, a boundary can be imposed as shown in Fig. 6.

(b) Keeping to the primitive case, the answer is that the system will endeavour to trade with the environment in the sense that some way of effecting the environment or some change of state in response to changes in the environment will elicit the reward of more energy.

(c) It will simplify the discussion to suppose that the system has its state changes coupled in some determinate manner to parameters of the environment and that the problem of getting a reward is thus a problem of sensing those changes in the environment which require

a particular response in order to achieve a reward. How, then, does the system appreciate its surroundings?

(d) It does so by developing specific sensory receptors. Note, first of all, that a thread structure is slightly sensitive to many disturbances, mechanical, chemical and electrical. Such disturbances, which will be encountered at the boundary, elicit some change of state. If this state change is unrewarded the disturbance in question will, however, be taken as irrelevant and will have little effect upon the system. But suppose that a disturbance, for example, a vibration or a change of acidity induces a state change which is rewarded (in other words suppose the environment is such that when part of the boundary is acid or vibrates some particular modification of the environment parameters makes more energy available to the system) then the system will adapt so that the boundary region becomes specifically sensitive to acidity or vibration. No teleological arguments are required to describe this process of building a sensory receptor for those variables which are sensed with advantage. Reward means permission to build more structures out of basic material. Thus a sensory receptor (which appears because the particular region of the boundary which *did* minimally respond to the environmental stimulus is duplicated and enlarged) will form as a logical consequence of specifically rewarding a system in which the elementary units have no well specified function and may not be regarded as components. Thread structures are just as good parts for pH meters and microphones as they are parts of memory registers and connections.

(e) From this it appears that the system can act like a natural historian and develop its own criteria of relevance and bring about any relation with respect to its environment.

(f) Although the input of such a system is badly specified, so far as a specialized observer is concerned this does not mean that the system is unable to distinguish input variables precisely. On the contrary, a system may discriminate variables by elaborating its receptor mechanisms to an arbitrary extent, if reward is contingent upon sensing the variables in question. However, it is possible to show by a recursive argument, that a specialized observer will nearly always be ignorant of the variables which are, at any moment, being sensed by the system.

(g) Even if the world of ferrous sulphate is finite the system can

trade in an indefinite number of ways with its surroundings and in this sense the environment boundary is a fiction. Using this assertion we can mark imaginary boundaries at arbitrary points in the system and examine the trade which takes place across them. This technique was tacitly adopted when discussing the different "mechanisms" for achieving "individuality." In other words different kinds of trading—different sensory receptors—are a special and dramatic case of the different mechanisms which exist as a commonplace feature of the system's activity.

(h) Again, these mechanisms evolve one from another. When we were discussing "individuality" the system "learned" about a primitive mechanism in order to evolve a less primitive and (to the natural historian) equivalent mechanism. At any stage in its development (when the system is observed over a short interval) there will exist an heirarchy of mechanisms, corresponding to different stages in its evolution.* Most of these will be vestigial, but the stability of the system derives from their existence and the possibility of their reactivation in adverse conditions.†

Rewarding Strategies

Returning to the strategy of a natural historian, it will be convenient to use some descriptive mathematics, again with the provision that the approach is tentative and will probably be replaced by more elegant and tractable techniques. In the first place let us recall:

- (a) That a "behavioral characteristic" is, for example, possession of "memory" or "docility" or "habituation."
- (b) The predictability condition "&" required by a natural historian.
- (c) The idea of a reference frame.
- (d) The assertion that a reference frame determines either:
 - (i) a set of relevant variables, or
 - (ii) a set of components which have some function—such as being neurones or something even more specific.

* This evolutionary structure is typical of biological systems, as pointed out by Professor Bishop.⁽²⁰⁾

† If, in the course of interaction with a system, we define certain regions as unit elements [Professor McCulloch's (Ref. 21)] reliability calculus is immediately applicable.

It will now be more convenient to adopt the latter restriction and to suppose that an observer in α sees components (the things which he calls neurones) as entities, $a_1, a_2, \dots$ in a network. Similarly, an observer in β sees $b_1, b_2, \dots$ (as the different entities he calls neurones). On the other hand, a natural historian takes different entities as being his basic elements on different occasions. He chooses the entities which allow him to make sense of and interact with the system.

We may now argue:

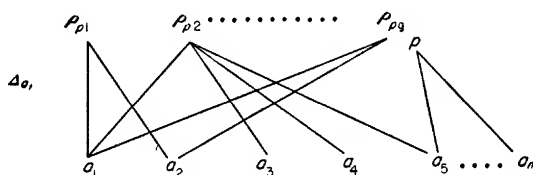
- (1) If $\psi_1 \psi_2 \dots \psi_e$ are the behavioral characteristics discussed by a natural historian in terms of the names $V_\mu V_\zeta$ of recognized states and the names $P_l P_m$ of stimulus procedures, there will be sets of names, some empty, such that all V_ζ and all P_l included in ξ_p refer to the p th-characteristic.
- (2) A Behavioral Characteristic may, at an instant t imply several distinct mechanisms. Thus, the p th-characteristic may imply any of g_p mechanisms.

$$\rho_1 \rho_2 \dots \rho_{g_p}$$

- (3) In a Reference Frame α we may assert, at t the existence of n_{α_t} active elements a and in β assert n_{β_t} active elements b . In general:

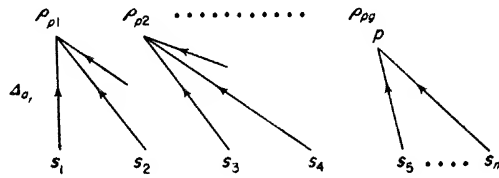
$$n_{\alpha_t} \neq n_{\beta_t}$$

- (4) There is a relation Δ which maps the $m_t < \sum_p g$ mechanisms active at t into the set of elements defined in α . Considering only the p th-characteristic this relation Δ_α may be:



If the mechanism p is in a self-organizing system then Δ is, as shown, a many to many relation. Further Δ is different for all $\alpha\beta$.

- (5) Sufficiently microscopic observation would discern elements $s_1 s_2 \dots s_{n_{0t}}$ with $n_{0t} > n_{\alpha t}$, $n_{0t} > n_{\beta t}$, for all α and β such that mapping Δ_0 is many to one like:



However, if the system is self-organizing, the elements s are inaccessible to a real observer.

- (6) Let $\Phi_{(p,d)t}$ equal the number of mapping arrows which converge upon $\rho_{(p,d)}$ in the mapping Δ_0 specified at t .
 (7) Let $\sigma_{(p,d)t}$ be a variable which is equal to 1 if and only if $\rho_{p,d}$ is active at t namely if and only if at least one mapping arrow converges upon $\rho_{(p,d)}$ in the mapping Δ_0 specified at t .
 (8) When a natural historian rewards a system he increases the value of a variable θ so that if a system is rewarded:

$$\theta_{t+1} > \theta_t$$

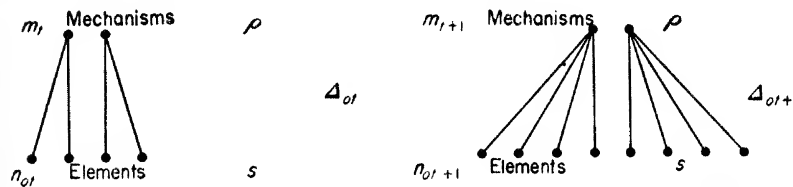
- (9) We have argued that the effect of increasing θ is to allow those mechanisms active at t to develop.

Thus $\Phi_{(p,d)t+1} > \Phi_{(p,d)t}$ if and only if $\theta_{t+1} > \theta_t$ and $\sigma_{(p,d)t} = 1$. In general this implies:

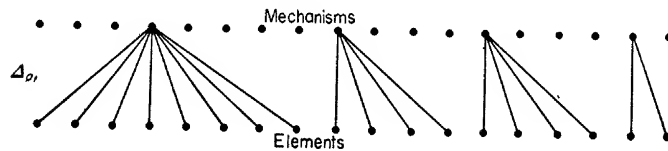
$$n_{0t+1} > n_{0t}$$

The effect of a reward upon Δ_α and Δ_β is, of course, unspecified.

- (10) Thus if a natural historian knows, or is able to determine that, the variable $\sigma_{p,d} = 1$ he can reward the system, so that $\rho_{p,d}$ becomes a dominant mechanism, for mediating the behavioral characteristic ψ_p . This will be the case only if the condition "&" is satisfied. Visually represented:



- (11) Consider only the behavioral characteristic ψ_p and thus assume that all V_μ and P_l are included in ξ_p .
- (12) Assume that each mechanism ρ_{pj} embodies a Composition Rule E_j in the sense that operations $V_\mu(E_j)P_l$ correspond to the operations of this mechanism.
- (13) If so the pair $E_j R_j$ will permit satisfaction of "&" and consequently, we write $E_j = E_j^*$. Equally, all Composition Rules E_i^* are embodied in some mechanism.
- (14) From (9) and (10) if the natural historian rewards the system only if the condition "&" is satisfied for his choice of E_i^* the mechanism ρ_{pi} will become dominant, the Composition Rule E_i^* will become more widely applicable, and b_i will increase.
- (15) The process is symmetrical for we can regard the natural historian as rewarded by achieving "&."
- (16) More specific strategies are needed if the system is being trained as a specific control mechanism. A single case will be suggested.
- (17) Let us apply these arguments to each p .
- (18) In this case it is possible to conceive a convergence of interaction by reward such that Δ_0 is so modified that for at least one α the mapping Δ_α is a many to one projection.



In this case, in its terminal condition, the system may be described in a single reference frame α which is not, however, determined at the outset of interaction. Thus the system will have been trained not to be a self-organizing system. It is a finite learning machine.

At the moment it is impossible to provide a general description of the way in which any specified objective should modify the process of interaction. One might avoid the problem by thinking of the natural historian as always a real person and the objective as

something he keeps in mind. This would, however, seriously restrict the admissible training procedures.

Moreover, a theoretical solution is almost certainly possible, since any arbitrary partitioning of a self-organizing system should produce two sub-systems, one of which is being trained by the other. As an Empirical Confirmation of this, it is possible to train a self-organizing system using an adaptive teaching machine^(22,23) (which is itself a finite learning machine) as a natural historian in place of the human being.

THE CHARACTER AND UTILITY OF SELF-ORGANIZING SYSTEMS

In conclusion let us review a number of self-organizing systems and consider how our knowledge of their natural history can be used.

By restricting the energy supply of an initially undifferentiated system (according to an appropriate rewarding strategy) it can be trained to act as a Control Mechanism. Clearly, however, this Control Mechanism has little in common with a programmed computer connected to a process by well defined input devices and output devices.

Take, for example, a chemical process. The control mechanism, in the present sense, is something which exists, perhaps on a catalytic surface, within the reaction vessel. It is sold by the cubic foot. Its inputs are sensory receptors, developed in the same manner as the rest of the network, and thus, although a variable such as pH may be sensed, we should not be able to indicate what part of the network sensed it, or to say that any part sensed it exclusively. The outputs of the control mechanisms might, in this case, be regional activations of the catalytic surface supporting the network.

Although metallic thread structures are conceptually useful, the particular model has a limited field of practical application. However, many building materials are available, and it seems likely that the optimum choice of materials will be different for different applications. Thus, in chemical control mechanism, the self-organizing system can sometimes be built up from the actual reactants.

There are two lines of thought, one leading us closer to conventional control mechanisms and one leading us further away. Pursuing the first, it is always possible to replace the process which

creates elements in a self-organizing system by an equivalent process which activates elements that already exist in large numbers. Similarly a network in which connectivity is built up is always equivalent to some network which has been derived from a large and fully connected plexus by removing connection.

Applying one or both of those equivalences we arrive at self-organizing systems either specially constructed, like some of the networks of the Illinois project,⁽²⁴⁾ or, when feasible, programmed on to a computer, like Selfridge's Pandemonium.⁽²⁵⁾ In the latter case, currency appears as an abstract cost function, but it is important to notice that reward must still mean permission to develop; for example, permission to take over more storage capacity in the computer or to replicate a demon. This connotation of reward is typical of a self-organizing system and distinguishes it from a superficially comparable structure.

Programmed systems of this kind are useful as research tools, as functional models of the brain, or when associated with a process by conventional input and output devices as control mechanisms which can deal with a non-stationary process.⁽²⁶⁾

Pursuing the second line of thought it occurs that nature has provided us with excellent physical models of self-organizing systems made out of protein. Beer⁽²⁷⁾ has discussed the possibility of using unicellulars, the colonial behavior of which is taken as organization, as the amplifying elements in a machine. He and I have examined models, where currency is food supply and unicellulars like paramecium, are active elements, sufficiently to show that such colonial organization may be coupled to a real process.

Self-organizing systems lie around us. There are quagmires, the fish in the sea, or intractable systems like clouds. Surely we can make these work things out for us, act as our control mechanisms, or perhaps most important of all, we can couple these seemingly uncontrollable entities together so that they control each other. Why not, for example, couple the traffic chaos in Chicago to the traffic chaos of New York in order to obtain an acceptably self-organizing whole? Why not associate individual brains to achieve a group intelligence?⁽²⁸⁾

These will remain intriguing ideas and no more until definite procedures are specified. There is a great deal of work to be done, but even at the present stage it is possible to envisage the form these

procedures must take. According to the present contentions there exist at least two sets of rules. The first set are rules, of a somewhat ephemeral character, which help a natural historian (or an adaptive machine) to interact efficiently with a self-organizing system (or in some cases determine the interaction of two self-organizing systems). As a result of the interaction some continually changing descriptive model is built up. Knowing this model, and in particular knowing the entities which are regarded at any moment as "elements" of the self-organizing system a second set of rules (which refer, for example, to reliability and habituation) come into play. The second set of rules are, in themselves determinate but they are applied to a "model" (the natural historian's model) which continually changes its relation to the real world.

REFERENCES

1. R. ASHBY, *Introduction to Cybernetics*, Chapman & Hall (1956).
2. G. PASK, Physical analogues for the growth of a concept, *Proc. Symposium, Mechanisation Thinking Process*, N.P.L. Teddington, 1958.
3. M. KOCHEN, Organised systems with discrete information transfer, *General Systems Society Yearbook*, 1957.
4. G. PASK, Organic control and the cybernetic method, *Cybernetica* 3 (1958).
5. W. McCULLOCH, Agathe Tyche, *Proc. Symposium, Mechanisation Thinking Process*, N.P.L. Teddington, 1958.
6. W. McCULLOCH, Reliability of biological systems. This volume, p. 264.
7. R. ASHBY, The mechanism of habituation, *Proc. Symposium, Mechanisation Thinking Process*, N.P.L. Teddington, 1958.
8. BUSH, MOSTELLER and THOMPSON, *A Formal Structure for Multiple Choice Situations, Decision Processes*, John Wiley, New York, 1954.
9. D. ROSENBLATT, On linear models and the graphs of Minkowski Leontief matrices, *Econometrica*, No. 2, 25 (1957).
10. R. ASHBY, *Cybernetica*, No. 2, Namur (1958).
11. G. PASK, Teaching machines, *Proc. 2nd Cong. Intern. Assoc. Cybernetics*, Namur, 1958.
12. H. von FOERSTER, Aspects in the design of biological computers, *2nd Cong. Intern. Assoc. Cybernetics*, Namur, 1958.
13. (For a recent discussion and full references to Prigogine's work.) C. FOSTER, A. RAPPORT and E. TRUCO, Some unsolved problems in the theory of non-isolated systems, *General Systems Society Yearbook*, 1957.
14. (For cellular catalysis.) J. POLONSKY, Quelques relexions sur l'aspect physique de la vie, *2nd Cong. Intern. Assoc. Cybernetics*, Namur, 1958.
15. H. von FOERSTER, Basic concepts of homeostasis, homeostatic mechanisms, *Brookhaven Symposia in Biology*, No. 10 (1957).
16. D. M. MCKAY, The epistemological problem for automata, *Automata Studies*, Princetown (1956).
17. Quarterly Report, Contract Nonr 1834(21), Electrical Engineering Research Laboratory, University of Illinois, Urbana (1959).

18. G. PASK, The growth process, *Proc. 2nd Cong. Intern. Assoc. Cybernetics*, Namur, 1958.
19. J. PLATT, Amplification aspects of biological response and mental activity, *Amer. Scientist*, Vol. 44, No. 2, April 1956.
20. G. H. BISHOP, The importance of feedback through the environment in the determination of motor activity. This volume, p. 122.
21. W. McCULLOCH, Agathe Tyche, *Proc. Symposium, Mechanisation Thinking Process*, N.P.L. Teddington, 1958.
22. G. PASK, Electronic teaching techniques, *British Communications and Electronics*, Vol. 4, No. 4, April 1957.
23. G. PASK, The teaching machine, *Overseas Engineer*, February 1955.
24. Reports of Contract Nonr 1834(21), Electrical Engineering Research Laboratory, University of Illinois, Urbana.
25. O. G. SELFRIDGE, Pandemonium, a paradigm of learning, *Proc. Symposium, Mechanisation Thinking Process*, N.P.L. Teddington, 1958.
26. S. BEER, The irrelevance of automation, *Proc. 2nd Cong. Intern. Assoc. Cybernetics*, Namur, 1958.
27. S. BEER, *Cybernetics and Industry*, English Universities Press (in print).
28. O. D. WELLS, Artorg Publications.

DISCUSSION

NEWELL: I am not quite sure that I understand this notion of the point of view on the natural historian. Let me ask you a question by proposing an example and seeing how this fits with your notion.

PASK: Very good.

NEWELL: If we wish to describe a human behaving, solving problems in logic, and we wish to build a program that describes him, then empirically it turns out about the only way any of us ever found out how to do it is to start from the beginning and proceed to build the program forward. We thus proceed for the first little bit of his behavior until that goes foul and then go back and proceed to reconstruct the program and again start out from the beginning until we run into the next little bit of behavior where it doesn't work, and so on, until in some sense we proceed through the entire structure of behavior.

Whenever you tell this to people they always ask why you don't in some sense first characterize what he does roughly and then specify this a little more and the empirical answer is, every time we try it this way we can't do it and the only time we ever seem to make progress is in some sense where we proceed almost at full detail at least the full details we can tolerate—and simply cruise in and follow, if I may now impose the word, the natural history, and follow in the footsteps of the person himself building the program as we go along.

Can you juxtapose this with your notion of a natural historian in some way?

PASK: If you will permit me, I will invert the example because the distinction I am making is only made for a finite observer. In other words, this distinction of a natural historian from a specialized observer is one that exists simply because observers are not almighty. If indeed they were, of course, they could get at a rock bottom, unambiguous, state description of the system. Equally well, they could split the system up into parts, each of which would have a definite function. However, we are imperfect observers but still find scattered around us in the real world, systems like dogs and elephants and such like things, which it would be too difficult, perhaps, too costly, for us to split down analytically in this way.

The question arises whether we can usefully observe these without trying to describe them in any "absolute" and "true" sense. Because although we are simply incapable of reaching the whole truth we can still usefully employ these admirable self-organizing systems.

The approach of a natural historian is a compromise, but I believe a very valid compromise. It is a way of dealing with systems that are not accessible to us and it is a method of doing it in much the manner which we use when we are trying to control another human being.

For example, you and I at the moment are having a conversation and you are putting an idea over and I am putting an idea over and we are using a procedure which relies in the first place upon inferring similarity between ourselves. We agree we are similar human beings. We understand each other's language and we suppose that certain similarity relations exist within the system as a whole.

I submit we infer all this on the basis of properties such as habituation, or such as having the kind of redundancy which Dr. McCulloch, I think, will describe this afternoon (that is the redundancy of potential command). Some of these properties, like the particular ones I have cited, may be expressed in rather exact terms. You can give a very good topological interpretation of the network structures that will evidence these properties.

On the other hand, some of them are vague. There are properties like "attitudes" and "mental set," and in such cases, we simply can't operate in terms that will allow us to express the concepts with which we infer similarity in mathematical language. The natural historian's approach is an endeavor to get a calculus which allows us to deal with situations like this and to interact with systems of this sort. A Natural Historian, as defined, has nothing to do with the *best* way of finding out *how* a system works. He is concerned actively to trade with the system. Given that he can, I further suggest it is possible to state, with precision, how we may interact with a system, even though we are unable to say how the system works. Does this answer you, sir, or have I missed your point?

NEWELL: That is a partial answer. I would like elucidation on this point. Is it fair to say that the human reacts by having a functional description rather than a human? Is it the kind of terms they use that work?

PASK: It is a functional description, sir, but in a changing language. You change your sample space and relevance criteria as you go along. At one stage of the conversation you talk about things which are logically speaking incomparable with other things introduced later. Yet you, yourself, and those in conversation with you, are able to tie these logically incomparable entities together.

CHAIRMAN MCCARTHY: Have you tried yet any of these other forms of stimuli such as pH?

PASK: We have made an ear and we have made a magnetic receptor. The ear can discriminate two frequencies, one of the order of fifty cycles per second and the other of the order of one hundred cycles per second. The "training" procedure takes approximately half a day and once having got the ability to recognize sound at all, the ability to recognize and discriminate two sounds comes more rapidly. I can't give anything more detailed than this qualitative assertion. The ear, incidentally, looks rather like an ear. It is a gap in the thread structure in which you have fibrils which resonate with the excitation frequency.

BASTIN (*King's College, Cambridge*): Concerning the production of a sense organ in a mechanical assembly, is this an illustrative model? If it is more can you tell us how it can help us discover things that we have not known before?

PASK: Yes, I think it is more than an illustrative model. The development of a sense organ is a dramatic way of showing what we mean by relevance criteria in connection with a machine.

Let us distinguish between that sort of machine that is made out of known bits and pieces, such as a computer (in terms of which we can make models of most physical assemblages, those normally studied in physics, and perhaps some of those we normally study in biology), and a machine which consists of a possibly unlimited number of components such that the function of these components is not defined beforehand. In other words, these "components" are simply "building material" which can be assembled in a variety of ways to make different entities. In particular the designer need not specify the set of possible entities.

Now, to say that this sort of machine will make sense organs is illustrative, because a sense organ is a rather special component construction in the sense that it specifies the boundary between the machine and its environment, and if the machine constructs its own sense organs out of the building material this boundary is apt to change continually. Further, it helps us to understand why we find it is difficult to observe these self-organizing systems, namely, because, looking at them in a single reference frame, in the capacity of scientific observers we cannot lay down definitions which allow us to compartmentalize the functional components of the machine or even the machine itself.

To see that this illustrative model is non-trivial you must recall that we reward the system without specifying, for example, that it will be rewarded for assembling component A and component B. We simply say it will get more reward—(meaning that it will be allowed to use more components) if a structure that acts as a sense organ is constructed. Clearly, unless you are prepared to take into account a variety of different ways of describing this machine (in different reference frames) you can't make sense of it, and that is what I meant by saying we must look at a self-organizing system as many coexisting models, mechanical, electrical and so on—rather than a single model—such as we have in a computer simulation.

Of course you can always transform the physical mechanism into an equivalent "single" model if you are landed with components such as those in a computer which have well defined functions, for example, a collection of storage devices. But you can do this only if you are prepared to have an indefinitely large number of these well defined components because you are led to construct a growing model which, if left on its own, will expand indefinitely. Of course, in the case of a computer model the idea of rewarding by a supply of energy or food is inappropriate. You have instead a value function so defined that it costs something to take over more bits of store.

This is certainly equivalent, but in fact, when we are thinking of real live systems which we are going to use, for example, in controlling chemical processes, we are really much more interested in the finite systems which are rendered non-bounded by the interesting condition that they can alter their own relevance criteria, and in particular, by the expedient of building sense organs, can alter their relationship to the environment according to whether or not a trial relationship is rewarded.

If I can have a minute to comment here, it is interesting that in any such system you will find a hierarchy of "vestigial" mechanisms of just the sort Professor Bishop commented upon yesterday. In other words, a primitive mechanism will evolve for sensing a variable and the machine will then learn about this mechanism and it will evolve a more sophisticated one. At any stage

of the system's development we should be able to observe both mechanisms more or less active. One is inclined to say it is crazy, that this machine has two ways of doing the same thing. But on second thought it is natural for a system to learn from its past attempts and make improvements.

THE RELIABILITY OF BIOLOGICAL SYSTEMS*

WARREN S. McCULLOCH

Institute of Technology, Cambridge, Mass.

I HAVE been working since 1917, in one way or another, on the problem of how man knows a number. For a while I wasted my time in psychology, then in physiology, and then in symbolic logic. Because I could not come to grips with the problem without better knowledge of the human brain, I studied medicine.

The most remarkable thing a doctor ever experiences is to deliver a baby. It comes out a certifiable suckling and it performs this sucking with great precision, although there are a few mistakes. This leads the physician to realize that, on the level of the human life, he is up against an enormously well-organized system that is consequently stable in its performances. So I look at the problem of self-organizing systems somewhat differently from anyone who begins with the presupposition of randomness.

The ultimate particles of physics must fit and stick to make atoms and those atoms must fit and stick to make molecules, and so it goes, up the scale of the natural objects, to the most complicated thing one knows about: man, always with enormously strong, close-range forces binding the parts together.

When we build a gadget out of components, we torture the materials; we lack atomic glues. We deal with large blocks of stuff sawed out like a piece of wood from a tree, where they belonged together, to make a thing, like this table, that couldn't withstand a hurricane for half an hour. The natural thing and the artifact are not alike.

* This work was supported in part by the U.S. Army (Signal Corps), the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U.S. Navy (Office of Naval Research); and in part by the National Institutes of Health.

I didn't appreciate how unlike they are even three or four years ago, for I thought of a nerve cell as a bag full of salt water. But now I am convinced that the inside of the nerve cell is as structured as the inside of a paramecium. I do not believe that even the water in it is free. There is no such disorder in it as there is in a solution.

Dr. von Foerster told us that order, whatever else it is, is fundamentally a redundancy of structure; so that which is ordered can be described in a smaller number of terms. That which is redundant is, to the extent that it is redundant, stable. It is therefore reliable. It is only out of redundancy that one can buy security. ✓

I went to M.I.T. to work on the circuit theory of brains because I was convinced that brains are devices for transmitting information. That is the way they have their truck with the totality of the world. Everything else is irrelevant. The physician needs to consider the energetics only when he has to treat an insane man or an epileptic; that is, only when the brain is diseased.

I thought that I would be able, in terms of information theory, to study the redundancies of code and the redundancies of channels, and so on, and simply transfer the proper theorems to the kind of redundancies that are most important in the bulk of the brain. As yet, we don't know all of them. There is redundancy of code and there is redundancy of channel; but there are kinds of stability you can buy with the redundancy of code that you cannot buy with the redundancy of channel, and vice versa. Personally, I have worked only on redundancy of calculation. By that I mean that the information is brought to a lot of so-called neurons, and these crummy neurons, working in parallel computation, can come out with the right answer even though the component neurons are misbehaving. I always expected, and so, I think, did Shannon and Elias, that we would be able to transfer many of our solutions from any one of these fields to the others; but they now appear to be radically different. The reliability you can buy with redundancy of calculation cannot be bought with redundancy of code or of channel. }

There is one more type of redundancy I want to speak about for a moment. That is the redundancy of potential command. Until two or three weeks ago I took it for granted that the redundancy of potential command was simply a redundancy of calculation. Now I am sure it isn't. Suppose you have, as you have in your reticular formation, or as you have in a naval fleet, knots of communication

in cells or ships and that each of these stations receives information from practically the whole of the organism or the fleet, but no two are coded alike. The information is probably pulse-interval-modulated and the signal probably carries the signature of its origin and word of what is going on there. Other stations may receive much the same information differently coded, but there is no assurance that all will get all of the incoming information.

In such a network as the reticular formation in your brain you have a structure in which any small number of cells can actually accumulate the necessary information, and, simply because they have that information, start buzzing. This decides that you attend to something to your right or go out and get your lunch, you go to sleep, or what not. If disease shoots out any particular cell, there are plenty of other cells that have much of the same information. A few of them, agreeing, run the works. Thus we have a redundancy of potential command in which knowledge constitutes authority. This needs study; but not discussion here.

What I shall present to-day is the redundancy of calculation. In it there are ridiculous troubles on account of misbehavior of small whole numbers that have delayed me unduly.

I have worked since 1952 to develop what I will call a probabilistic logic in the sense in which John von Neumann used the phrase. By "probabilistic logic" von Neumann did not mean a logical calculus composed of certain functions of arguments which were uncertain—i.e. a calculus of probabilities. [He meant a calculus in which the functions were uncertain, whether the arguments were certain or only probable.] This, his usage, has nothing to do with multiple truth-valued logics, which I have never found to be of any help in this problem. The neuronal diagrams that he and we have employed are formal devices for arithmetical operations, and their resemblances to real neurons is of no logical significance and only a stimulus to neurophysiological investigation. The neuronal nets are not a development of Birkhoff's lattice theory, and, though they have superficial resemblances, neither has assimilated the other. The operations of these probabilistic symbols upon each other are not simple matrix multiplications.

I have brought you uncorrected copies of my Quarterly Progress Report on probabilistic logic. It constitutes Part B of my presentation. I have also brought you a very rough first draft of the paper I am

about to read to you. It will appear in a later Quarterly Progress Report. Call it Part A.

I would like to present the results of half a dozen years of thinking in half an hour. To make it easier for you I have put at your disposal uncorrected copies of both parts of what I am about to say. Both are but continuations of work presented in Teddington under the title, *Agathe Tyche*.⁽¹⁾ There I described solutions to these problems raised by John von Neumann,^(2,3) as to the action of the nervous system. They are concerned with (1) circuits which have the same output for the same input when all component neurons undergo a common shift of threshold so that each component computes some new function of its input; (2) the design of circuits more reliable than their components; and (3) the nature of neurons that makes them more flexible in computation than any man-made devices.

The symbols used there, and here, to describe the functions computed by neurons are derived from Venn's⁽⁴⁾ diagrams for the intersection of classes. In this case, the classes are composed of events each of which is the signal that an event of that class has occurred. When a jot appears in such a diagram in any place it indicates that the neuron whose action it represents is supposed to fire under that condition. Thus each Venn symbol pictures Wittgenstein's Truth Table⁽⁵⁾ for that function. Every line in a Venn diagram divides all other spaces into two spaces. Thus the Venn symbol for a neuron with δ inputs has 2^δ spaces. Here are my diagrams (Fig. 1).

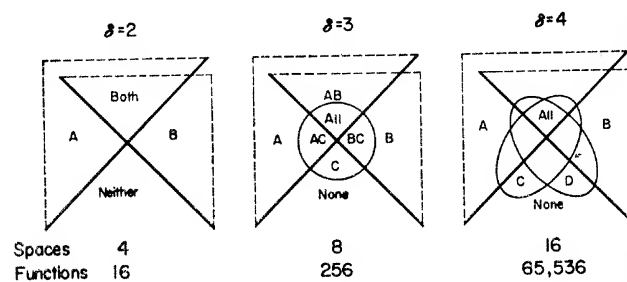


FIG. 1. Venn symbols for δ equal to 2, 3 and 4.

For larger δ it is more convenient to use Oliver Selfridge's device consisting of cosine waves each twice the amplitude and half the

frequency of its predecessor, so as to generate the sequences of spaces S_0 through $S_{2^{\delta}-1}$ of Table I of part B. It is only of existential importance in part A, for it allows us to consider δ of any size, though the Venn symbols actually exemplified in it are for $\delta = 3$.

PART A

INFALLIBLE NETWORK OF FALLIBLE NEURONS

When I presented *Agathe Tyche* I thought that no circuit could preserve error-free action if it were composed of 3 neurons each receiving only two inputs. To correct this, let me replace the jot in any Venn symbol by 1 if the jot is always present, by 0 if it is always absent, and by p if it is present with a probability p due to a shift of the threshold θ of the neuron it represents. Thus



represents a neuron which always fires when A alone or both A and B occur, and fires with a probability p when B alone or neither

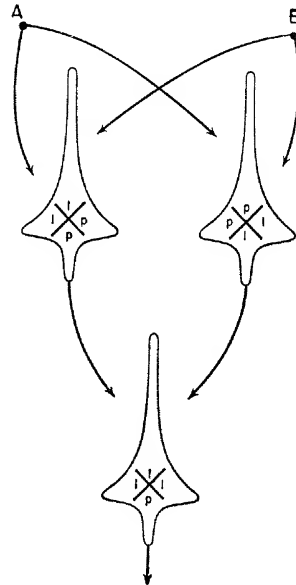


FIG. 2. An infallible net of $\delta = 2$ neurons.

A nor B occurs. For error-free operation, some spaces in the Venn symbols must have 0 or 1. I place each symbol inside the iodogram of its neuron and draw Fig. 2 in which each of those of the first rank receives signals from A and from B , and both play on the output neuron at the bottom. A formula of the first rank is one whose arguments are independent. A formula of the second rank is made by replacing the arguments with formulas of the first rank. Thus in Fig. 2, the top two neurons are of the first rank and the lower is of the second rank.

For nets of neurons of 2 inputs, i.e. whose δ is 2, zero error can be achieved only for tautology and contradiction with some p in every chi (χ), that is, by keeping the fixed jots or blanks in such positions that, when added, they form tautology or contradiction. But, as the complement of tautology is contradiction and that of contradiction is tautology, these lead only to themselves and to no significant functions of their primitive propositions. I might write this

$$\left(\begin{array}{c} | \\ \diagup \quad \diagdown \\ | \quad p \\ \diagdown \quad \diagup \\ p \end{array} \right) \quad \begin{array}{c} | \\ \diagup \quad \diagdown \\ | \quad | \\ \diagdown \quad \diagup \\ p \end{array} \quad \left(\begin{array}{c} p \\ \diagup \quad \diagdown \\ p \quad | \\ \diagdown \quad \diagup \\ | \end{array} \right) \equiv \left[\begin{array}{c} | \\ \diagup \quad \diagdown \\ | \quad | \\ \diagdown \quad \diagup \\ | \end{array} \right]$$

Thus this circuit always computes tautology, for it fires for any combination of the truth values of A and B . Similarly I may write for contradiction

$$\left(\begin{array}{c} p \\ \diagup \quad \diagdown \\ | \quad p \\ \diagdown \quad \diagup \\ p \end{array} \right) \quad \begin{array}{c} o \\ \diagup \quad \diagdown \\ o \quad o \\ \diagdown \quad \diagup \\ p \end{array} \quad \left(\begin{array}{c} | \\ \diagup \quad \diagdown \\ p \quad | \\ \diagdown \quad \diagup \\ | \end{array} \right) \equiv \left[\begin{array}{c} o \\ \diagup \quad \diagdown \\ o \quad o \\ \diagdown \quad \diagup \\ o \end{array} \right]$$

The two neurons of the first rank need only be such that the ones in each added to those in the other are sufficient to insure a one in every place—i.e. they must be complements under tautology. Yet, since the complement of tautology is contradiction and, of contradiction, is tautology, these are useless for generating any other function free of error. Tautology and contradiction tell us nothing about the world, and it is only these other functions that

are significant. I should like to add here that if one is given a single significant function without error one can compute it and its complement without error, using only fallible neurons.

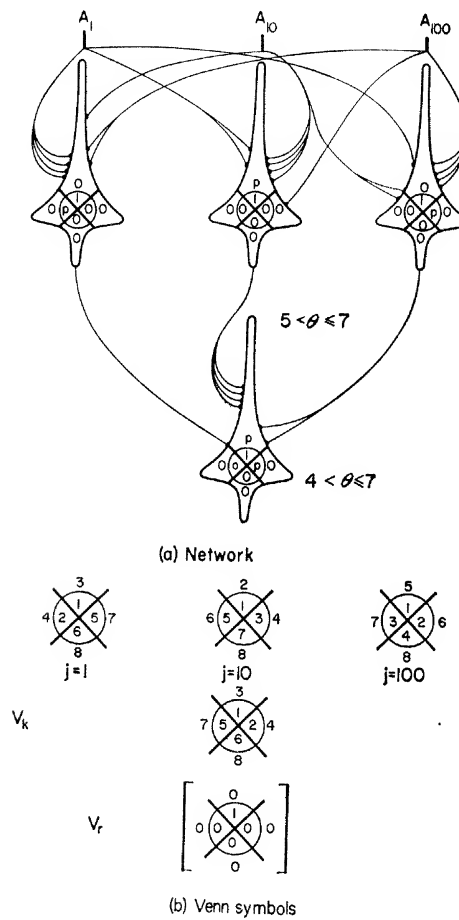


FIG. 3. Error-free net of $\delta = 3$ neurons.

This limitation disappears with $\delta = 3$. For instance, consider the net of Fig. 3(a) for nondegenerate* $\delta = 3$ neurons. In Fig. 3(b)

* A nondegenerate neuron is a neuron for which a change of one in θ either increases or decreases the number of jots in the Venn symbol by one.

are shown the Venn symbols for these neurons with the order in which jots are added as θ decreases indicated—i.e. if $\theta = 7$, the symbol has a jot in position 1; if $\theta = 6$, jots are located in 1 and 2; if $\theta = 5$ they are located at 1, 2, and 3; and so on, until for $\theta = 0$ all eight spaces in the symbol contain jots. The neurons of the first rank will all be certainly fired when all three inputs are present. The first may also be fired by the combination of inputs A_1 and A_{100} ; the second may be fired by the combination of inputs A_1 and A_{10} ; and the third by A_{10} and A_{100} . Thus none of the first rank neurons can be fired by a single input, and furthermore no combination of two inputs can fire more than one. The output neuron will certainly be fired when all three first rank neurons are fired and may be fired by any combination of two of them. However, it has been shown that only the presence of all three inputs A_1 , A_{10} and A_{100} will cause more than one of the first rank neurons to fire. Therefore only this combination will result in an output from the net. Thus in each Venn symbol any of the p 's can be replaced by a 0 or 1 without affecting the network output of the network.

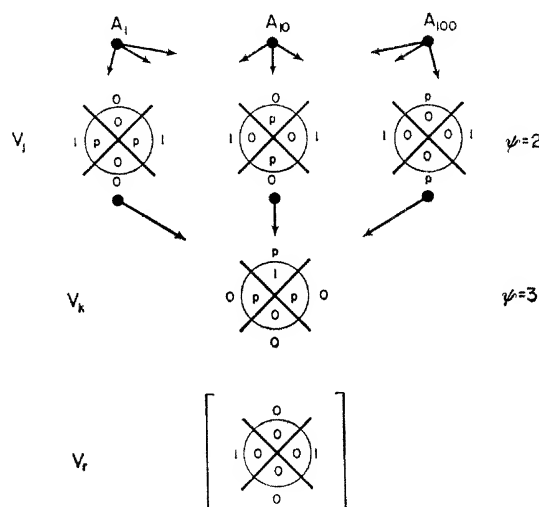


FIG. 4. Derivation of an infallible net.

The following algorithm will serve for the derivation of error-free nets when δ is odd. Let $J = 2^\delta \bmod \delta$, and let ψ be the number of

jots that may be added harmlessly to the Venn symbol that can stand fewest additions of any Venn symbol in its rank; write the desired function (V_j) of J or fewer jots as 1's in more than half of the V_j . In V_k enter the required 1 in the proper space, and p in all of its other spaces for the intersections of more than half of its arguments (see Fig. 4). There are $2^{\delta-1}$ of these spaces, one of which contains the 1: Therefore the number of p 's in V_k , call it ψ_k , is $2^{(\delta-1)} - 1$.

In every V_j there remain $2^\delta - J$ empty spaces in which p 's may be placed if they do not occur in the same space in more than $\frac{1}{2}(\delta - 1)$ of the V_j . This permits them harmlessly in $\frac{1}{2}(2^\delta - J)(\delta - 1)\delta^{-1}$

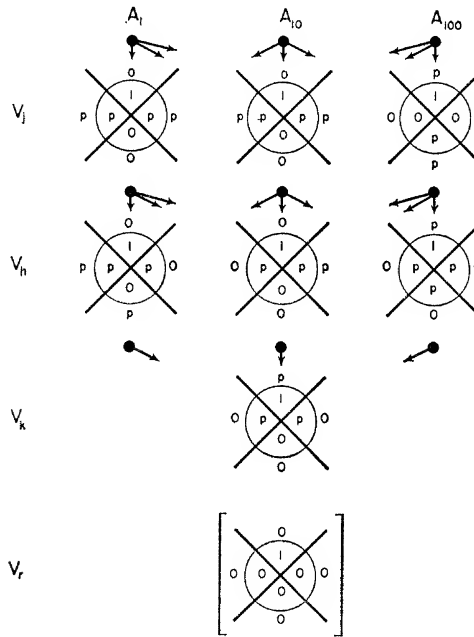


FIG. 5. Infallible network with third-rank output neuron.

spaces of each V_j . Now, for δ odd, J is 2 if δ is prime. Harmless p 's are most numerous when J is smallest. Hence the maximum $\psi_j = (2^\delta - 2)(\delta^{-1})\frac{1}{2}(\delta - 1) = (2^{\delta-1} - 1)(\delta - 1)\delta^{-1}$.

Writing ψ^* for $2^{-\delta}\psi_{\max}$, we have

$$\psi_k^* = 2^{-\delta}(2^{\delta-1} - 1) = \frac{1}{2}(1 - 2^{1-\delta});$$

and

$$\psi_j^* = 2^{-\delta}\delta^{-1}(2^{\delta-1} - 1)(\delta - 1) = \frac{1}{2}(1 - 2^{1-\delta})(1 - \delta^{-1})$$

Hence,

$$\lim_{\delta \rightarrow \infty} \psi_k^* = \lim_{\delta \rightarrow \infty} \psi_j^* = \frac{1}{2}$$

The binomial coefficient for the central term when δ is even has to be allotted either to the V_j or else to V_k , and its effect diminishes comparatively slowly, without affecting the limits.

Functions of J or fewer 0's can be constructed similarly. Any additional 1 or 0 after J decreases ψ_j by 1, reaching a maximum loss of $\frac{1}{2}(2^{\delta} - J)$ in the V_j , thus effectively halving ψ_j^* . When not all J are used for 1's or 0's, this space can be occupied in only some of the V_j and so ψ_j^* is not increased.

For nets with 2 ranks, V_j and V_h neurons, and one third-rank output neuron V_k , ψ_j and ψ_h increase to equal ψ_k for all neurons, and almost all may have one additional p harmlessly as in Fig. 5, where $\psi = 3$ in $V_{j=100}$ and in $V_{h=10}$ and in V_k .

We compute the number of nondegenerate infallible nets of two ranks thus: For each V_j there are

$$\psi_j! \left(\psi_j \frac{\delta + 1}{\delta - 1} \right)!$$

which meet the requirements, and for V_k there are

$$\psi_k! \left(\psi_k \frac{\delta + 1}{\delta - 1} \right)!$$

The number of nondegenerate nets is

$$\psi_j! \left(\psi_j \frac{\delta + 1}{\delta - 1} \right)! \psi_k! \left(\psi_k \frac{\delta + 1}{\delta - 1} \right)!,$$

second we define the least as $x + 1$ and suppose that they differ from each other by 1 step in θ . Then

$$\frac{\delta + 1}{2}x + \sum_1^{\frac{1}{2}(\delta + 1)} (\text{integers}) > \frac{\delta - 1}{2}x + \sum_{\frac{1}{2}(\delta + 1) + 1}^{\delta} (\text{integers})$$

and

$$x > \frac{\delta + 1}{2} \left[(\delta + 1) - \left(\frac{\delta + 1}{2} + 1 \right) \right] - \frac{\delta + 1}{2},$$

which is
$$\frac{\delta + 1}{2} \cdot \frac{\delta - 3}{2}$$

Hence the least term is

$$\left(\frac{\delta + 1}{2} \cdot \frac{\delta - 3}{2} \right) + 2 = \left(\frac{\delta - 1}{2} \right)^2 + 1,$$

and the greatest is $[(\delta + 1)/2]^2$; but this, which yields ψ_k^* , is as far from equal strengths as possible. The allowable change in threshold is decreased only by a factor $1/2\delta[(\delta^2 - 1)/(\delta^2 + 3)]$.

For example: if V_k has 7 inputs, each with a value of +13, its output is error-free for $39 < \theta < 91$, which is reduced by letting the inputs range from +10 to +16 to $45 < \theta < 91$. Thus the remaining usable range of θ , $\Delta^*\theta$, exceeds $\frac{1}{2}$, while $\psi_k^* = [\frac{1}{2} - (1/2^7)]$.

Note that if one is willing to forego the possibility of a jot appearing in the position for none in V_k , thus reducing ψ_k to $[\frac{1}{2} - (1/2^6)]$, the value of $\Delta^*\theta$ is 66.6% for equal afferents and 56% for afferents ranging from +10 to +16.

Since the measured variation, $\Delta\theta$, of real neurons is $\pm 5\%$, we look next at the permissible independent variation of the strength of afferent signals to V_k , when these strengths were intended to be equal. Clearly, any selection of $(\delta - 1)/2$ signals has a maximum sum less than the minimum sum for δ signals. Let t be the intended strength of a and, for large δ , it is nearly $(2^{\delta-1})^4$, which is a large number but, divided by the number of all nets, is the negligible fraction

$$\frac{(2^{\delta-1})^4}{(2^\delta!)^{\delta+1}}$$

If nets with more jots than J were equally numerous, and they are not, it would not multiply this fraction by more than $(2^{2^\delta} - 4)$ which still leaves it negligible. Chance is unlikely to produce such nets or to discover them among nets supposed equiprobable.

No comparable measures of the number or fraction of diagrams that are degenerate can be similarly computed, either because (1)

they do not change the function computed with every step of θ or because (2) they change the symbol by two or more jots per step in θ . What misled me in *Agathe Tyche* was that I had examined only output neuronal diagrams in which $\delta \leq 3$ and in which there was no inhibition of afferents by afferents, but only a direct action on the recipient neuron, and that, among these, only degenerate diagrams give ψ_k . Clearly, the degeneracy of the second kind is greatest when all afferents have equal signal strength, and of the first kind when the strength of the afferents goes up maximally between none and the least one, etc. The first kind is therefore minimized by using the smallest whole numbers possible, and the second by having them as unequal as possible.

Since in V_k the spaces for $(\delta + 1)/2$ arguments must be filled before any spaces for fewer arguments are filled, inputs each equal to one would produce the desired result with no degeneracy of the first kind but with maximum degeneracy of the second kind. To minimize the signal and Δt its variation. Then

$$\sum_0^{\frac{1}{2}(\delta-1)} (t + \Delta t) < \sum_0^{\delta} (t - \Delta t)$$

Hence

$$\frac{|\Delta t|}{t} < \frac{\delta + 1}{3\delta - 1} = \frac{1}{3} \left(1 + \frac{4}{3\delta - 1} \right)$$

which always exceeds $\frac{1}{3}$. For example, for $\delta = 3$, with the intended strength, t , equal to 2, we have, $1 < (t \pm \Delta t) < 3$. To retain $\Delta\theta$ of $\pm 5\%$ of its intended value reduces these limits to

$$1.05 < (t \pm \Delta t) < 2.85, \text{ or } |\Delta t|/t = 0.425.$$

[When a variation of this size is experimentally produced in the afferent termination in the spinal cord of the cat, it alters the circuit action. Even post-tetanic potentiation and convulsive doses of strychnine alter the voltage of signals by less than 10%].

If we hold $\Delta\theta$ to a minimum and the signals to their intended strength, we may permit variation, Δs , in the connections, or synapsis, s , to obtain a similar limit.

$$\frac{|\Delta s|}{s} < \frac{1}{3} \left(1 + \frac{4}{3\delta - 1} \right)$$

But it is more reasonable to suppose that, on real neurons, the sum of the afferents, $\sum_1^\delta (s \pm \Delta s) = \delta s$; whence

$$|\Delta s|/s = \delta + 1/\delta - 1$$

whose limit as $\delta \rightarrow \infty$ is unity. For $\delta = 3$, to the nearest integer, $s \pm \Delta s = 2 \pm 3$ (see V_k of Fig. 6) the range of θ is from (-1) to $(+7)$ and $5 \leq \theta \leq 6$, or $\Delta\theta = \pm 6.25\%$.

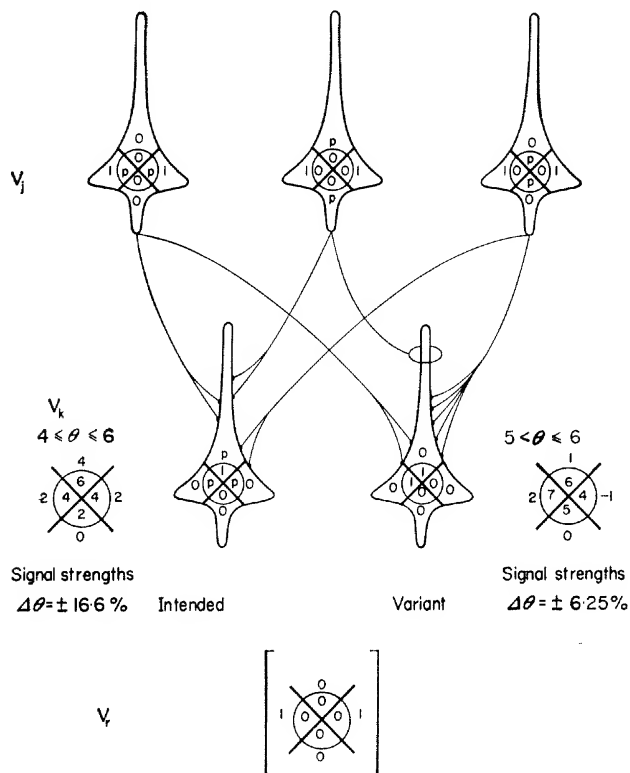


FIG. 6. Network illustrating variation in synapses.

From the preceding paragraphs it is clear that the number of p 's that may appear in any Venn symbol for a given change in θ is fixed only for nondegenerate diagrams, and that, for the degenerate diagrams, the fractional change in the jots is generally less than the fractional change in θ . Thus the actual reliability tends to exceed

expectation as the specifications of strengths of signals or of coupling (i.e. synapsis) are randomly perturbed. But it would be a work of supererogation to inquire into this in general, for the nature of real or of artificial neurons and the statistical specifications of their connections necessarily determine the weight to be allotted to each factor.

If an educated guess as to real neurons and their nets is now permissible, it must take the general form of a $\Delta\theta$ of $\pm 5\%$, a Δt of $\pm 10\%$, which leaves Δs , for $\delta \geq 3$, less than $\pm 100\%$ for synapsis for neurons of the second or higher rank. This presupposes that neurons of the first rank are relatively closely specified in synapsis in order to segregate possible errors. This is in harmony with much that is known of the auditory system, wherein pitch, loudness, and direction are initially decoded and thence transmitted over separate channels or in dissimilar codes. It begins to look as though the same were true of vision, of proprioception and of the stages of afferents from the skin following detection and amplification. Thus by the time information from any source reaches our great central computers we are in a region wherein crude specifications of statistical kinds should insure error-free calculation despite gross perturbation of threshold, of excitation, and even of local synapsis. This conclusion follows from two assumptions: first, that we are dealing with a parallel computer of more than two afferents per neuron, and second, that the functions that their neurons compute are sufficiently dissimilar to insure, at at least one level, incompatibility of error in the functions computed. All else may be safely left in large measure to chance.

PART B

ON PROBABILISTIC LOGIC

Formulae for any logical function of two arguments, and these are the usual ones of symbolic logic, can be reduced to a single Venn symbol by the rules given in *Agathe Tyche* for operators with jots in x 's. It is a special case of what we state here for any number of inputs and for the likelihood, $0 \leq p_i \leq 1$, of a jot appearing in the appropriate space, S_j , of the Venn symbol.

Any statement of the finite calculus of propositions can be expressed as follows. Subscript the symbol for the δ primitive propositions A_j , with j taking the ascending powers of 2 from 2^0 to

$2^{\delta-1}$ written in binary numbers, A_1, A_{10}, A_{100} , etc. Construct a V table with spaces S_i subscripted with the integers i , in binary form, from 0 to $2^{\delta} - 1$ (cf. columns S_i of Table 1).

TABLE 1

Arguments	Spaces*	First Rank Output			Input-Output
A_j	S_i	$V_{h=1}$	$V_{h=10}$	V_k	V_r
None	S_0	p_0	p_0	p'_0	p''_0
A_1	S_1	p_1	p_1	p'_1	p''_1
A_{10}	S_{10}	p_{10}	p_{10}	p'_{10}	p''_{10}
$A_{10} \quad A_1$	S_{11}	p_{11}	p_{11}	p'_{11}	p''_{11}
A_{100}	S_{100}	p_{100}	p_{100}		p''_{100}
$A_{100} \quad A_1$	S_{101}	p_{101}	p_{101}		p''_{101}
$A_{100} \quad A_{10}$	S_{110}	p_{110}	p_{110}		p''_{110}
$A_{100} \quad A_{10} \quad A_1$	S_{111}	p_{111}	p_{111}		p''_{111}

EXAMPLE 1

Let V_h be $\delta = 3$, V_k be $\delta = 2$ (as in Table 1), then we have:

$$\begin{aligned}
 p''_0 &= p'_0 & q_{0,1} & q_{0,10} & p_1 &= p_0 & q_{1,1} & q_{1,10} \\
 &+ p'_1 & p_{0,1} & q_{0,10} & &+ p_1 & p_{1,1} & q_{1,10} \\
 &+ p'_{10} & q_{0,1} & p_{0,10} & &+ p_{10} & q_{1,1} & p_{1,10} \\
 &+ p'_{11} & p_{0,1} & p_{0,10} & &+ p_{11} & p_{1,1} & p_{1,10} \\
 \\
 p''_{10} &= p'_0 & q_{10,1} & q_{1,10} & p''_{111} &= p'_0 & q_{111,1} & q_{111,10} \\
 &+ p'_1 & p_{10,1} & q_{1,10} & &+ p_1 & p_{111,1} & q_{111,10} \\
 &+ p'_{10} & q_{10,1} & p_{1,10} & &+ p_{10} & q_{111,1} & p_{111,10} \\
 &+ p'_{11} & p_{10,1} & p_{1,10} & &+ p_{11} & p_{111,1} & p_{111,10}
 \end{aligned}$$

* For this Venn diagram I am indebted to O. Selfridge and M. L. Minsky.

Each i is the sum of one and only one selection of j 's and so identifies its space as the concurrence of those arguments ranging from S_0 for none of the A_j to S_{2^s-1} for all of them (cf. column A_j of Table 1). Thus 1 and 100 are in 101 and 10 is not, for which we write $1 \in 101$ and $100 \in 101$ and $10 \notin 101$. In the given logical text first replace A_j by V_h with a 1 in S_i if $j \in i$ and with a 0 if $j \notin i$, which makes V_h the truth-table of A_j with $T = 1$, $F = 0$.

Repeated applications of a single rule serve to reduce symbols for probabilistic functions of any δA_j to a single table of probabilities, and similarly any uncertain functions of these, etc. to a single V_r with the same subscripts of S_r as the V_h for the A_j (cf. V_r of Table 1 where $q = 1 - p$). This rule reads: replace the symbol for a function by V_k in which the k of S_k are again the integers in binary form but refer to the h of V_h , and the p_k of V_k betoken the likelihood of a 1 in S_k . Construct V_r and insert in S_r the likelihood p_r of a 1 in S_r computed by equation (1). (See example 1.)

$$p_r = \sum_{k=0}^{k=k_{\max}} r'_k \prod_{\substack{h \in k \\ r=i}} p_{ih} \prod_{\substack{h \notin k \\ r=i}} (1 - p_{ih}) \quad (1)$$

REFERENCES

1. W. S. McCulloch, Agathe Tyche: of nervous nets—the lucky reckoners, *Proc. Symp. on Mechanization of Thought Processes*, N.P.L., Teddington, England.
2. J. von NEUMANN, Probabilistic logics and the synthesis of reliable organisms from unreliable components, *Automata Studies* (ed. by C. E. SHANNON and J. MCCARTHY), Princeton University Press, Princeton, N.J. (1956). ✓
3. J. von NEUMANN, *Fixed and Random Logical Patterns and the Problem of Reliability*, American Psychiatric Association, Atlantic City, 12th May 1955. ✓
4. J. VENN, *Phil. Mag.*, July 1880.
5. L. WITTGENSTEIN, Logisch-philosophische Abhandlung, *Ann. der Naturphil.* (1921).

DISCUSSION

UTTLEY: I have thought for a long time about this fascinating theme of reliability out of unreliable units but there is something that is not physiologically clear that worries me very much. Dr. McCulloch started by telling us about the incredible stability and organization of living systems as opposed to desks and chairs and he also said that he once thought a neuron was a bag with salt water inside of it and he now thinks it is highly organized inside.

McCULLOCH: I certainly do.

T

UTTLEY: Well, now, I feel that neurons are not crummy, because to suggest that neurons are crummy is rather like suggesting they are filled with salt water. When we first put a microelectrode into some high-level part of the cortex we see apparently random firing impulses. I think it would be very brave to suggest that there is some true random principles going on here. We may use the probability theory to sidestep our ignorance, but I am worried about their being real unreliability about. We know in the rods in the retina there is pretty certainly some true randomness arising from the quantum uncertainty. What I rather doubt is that some future Heisenberg will come along and tell us that in the neuron, where we really do know all we can know about it, there is still some uncertainty. So I am very worried about the apparently different attitudes you have about biological systems being incredibly redundant and organized and yet crummy.

MCCULLOCH: Well, let's begin. We are interested primarily in the nervous system. Half of all the people that are in the hospitals of the United States are there because something is wrong in the nervous system; so they constitute a very large fraction of all our woes.

That these things go wrong all too often any psychiatrist can tell you. When you have as your daily customers the man that says that his automobile is twenty-two and twenty-two is his house number and goes to bed in the automobile, you have to realize his circuit doesn't always work correctly. (Laughter.)

Suppose you have to design a circuit to last a lifetime without replacing components. I am 60 years old. In my cerebellum there must be by now at least a 10% loss in the Purkinje-cell population—perhaps a 20% loss. Yet the circuit still works. You have to design a brain so it can get drunk and still find its way home. You have to build it so you can anesthetize it without its dying of respiratory failure. It must take all kinds of abuse, general shifts of threshold, local shifts, etc.

I was amazed when we ran into a 10% variation in the striking voltage at the node of Ranvier, which is probably a fair model of the trigger-point. I had guessed from old measurements of the size of the node of Ranvier that the variation would be less than one-tenth of 1%, for it seemed larger compared to thermal noise at body temperature. At Teddington, where Dr. Young was in the Chair, I asked him whether I was crazy or whether the measurements were crazy. He said that from electron-microscope studies he believes the effective surface of the node of Ranvier is extremely small.

When we attempt to miniaturize to the extent to which brains are miniaturized, we are going to have trouble at the levels of the component. We have to make use of a redundancy of computation to secure reliability of unreliable components.

That is the crucial thing as far as I can see it. We can pack more and more computing surfaces into the volume but then we have to put these smaller components together so as to build back the stability we would have if we had one large surface.

NEWELL: On the potential command, I was strikingly reminded of the pandemonium model. I wondered if there was some similarity here. And I was also reminded of the high rigidity in all the programming languages that we have to date which are rigid sequential languages. Pandemonium on the other hand is a rule of which it might be said, instead of doing whatever process comes next, you will always do what process shouts loudest. And the principle I thought you were enunciating is the principle that the man who has the information stands up and says, "I know"—in fact, he is the only one who stands up. We have had one or two other examples of this mostly in the language translation

field in which one builds a programming language where one simply salts away in the expression large numbers of programs and then the one with the highest priority says, "I am supposed to work." He works, and then the next guy says, "I am supposed to work," and he works.

I think there is a very significant trend in the programming language field to really get new ways of dealing with the language other than sequential controls.

McCULLOCH: I think it is crucial. I agree with you entirely. I am not sure I caught the totality of the first part of your statement.

NEWELL: It is very like pandemonium.

McCULLOCH: Yes, it is very like pandemonium in many ways and it will be more like pandemonium when they install it in Jerry Lettvin's artificial neurons. These have the full Hodgkin-Huxley eighth order, nonlinear differential equation to describe their behavior.

COMPUTATION, BEHAVIOR AND STRUCTURE IN FIXED AND GROWING AUTOMATA*†

ARTHUR W. BURKS

University of Michigan, Ann Arbor, Michigan

1. INTRODUCTION

THIS paper is concerned exclusively with deterministic automata. Most of this symposium has been devoted to probabilistic automata, so a word of explanation is in order.

It is perhaps true that deterministic theories are inadequate to account fully for growth, learning, mutation, evolution, and the actual operation of digital computers, and hence that a probabilistic theory is required for a full explanation of all these phenomena. Nevertheless, a discussion of deterministic automata is not out of place in a symposium on self-organizing systems. We cannot fully understand probabilistic automata until we know the limits of deterministic ones, i.e. what deterministic automata cannot do. Moreover, a probabilistic automaton may be regarded as a deterministic automaton to which has been added a probability measure governing the transitions between states. Just as both deterministic and probabilistic theories have been important in physical science, we may expect both kinds to be fruitful in the study of automata.

A complete theory of self-organizing systems must show exactly what role determinism plays in these systems. I accept von Neumann's suggestion that automata theory must include probabilistic as well as deterministic logics,⁽¹⁾ but I also believe with him that

* This research was supported by the Office of Naval Research [Contract 1224(21)].

† John Holland and Jesse Wright have made many helpful suggestions.

deterministic analyses of such phenomena as self-reproduction are of interest (see Section 4 below).

Various species of deterministic automata have been studied. Turing machines have been intensively investigated, both directly and through the medium of recursive functions.⁽²⁾ Fixed (finite) automata have received much attention of late, and von Neumann worked with self-reproducing automata. Little has been done, however, to relate these separate inquiries. In the present paper I will make some suggestions directed toward a unified theory of deterministic automata.

2. STRUCTURE, BEHAVIOR AND COMPUTATION IN FIXED AUTOMATA AND GENERALIZED TURING MACHINES

I will begin with fixed automata and Turing machines, and discuss later self-reproducing machines and growing automata generally. A fixed automaton consists of a finite number of switch and delay elements interconnected in a pattern or structure which is the same at every moment of time.^(3,4) Figure 2.1 represents a very simple

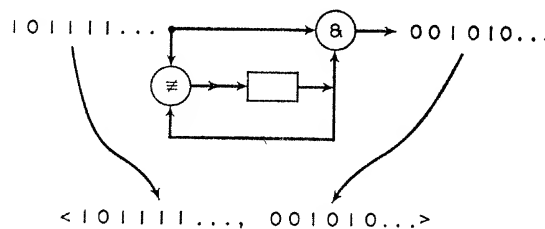


FIG. 2.1. Behavior of fixed automaton.

fixed automaton which deletes every other pulse (1) that is fed into the input. A digital computer with a finite tape is, of course, a fixed automaton.

The structure of a fixed automaton is to be distinguished from its behavior. The *structure* consists in the fixed pattern of interconnections of the elements. The structures of two automata are the same if the nodes (i.e. the junctions of the wires of the elements) of these automata can be placed in one-one correspondence so that corresponding nodes are nodes of identical elements. Thus if we

make the conjunction of Fig. 2.1 into a disjunction we obtain a different structure.

The behavior of an automaton consists, roughly speaking, of the relation of input stimuli to output responses. A more precise definition of behavior may be given in terms of temporal sequences of states, where the relevant times are the non-negative integers 0, 1, 2, 3. . . . Let us call an infinite temporal sequence of input states to an automaton an *input history* and an infinite temporal sequence of output states an *output history*. A *behavior pair* of an automaton consists of an input history paired with the output history which is produced when that input history is impressed on the inputs of the automaton. A behavior pair for the automaton of Fig. 2.1 is shown in the lower part of the figure. The *behavior* of a fixed automaton is the set of all behavior pairs of that automaton.

We turn next to Turing machines. As Turing originally conceived them each is essentially a fixed automaton connected to a single infinite tape. The fixed automaton scans one square of the tape at any one time, and can move the tape (or move relative to the tape) one square at a time. We will generalize Turing's concept in a

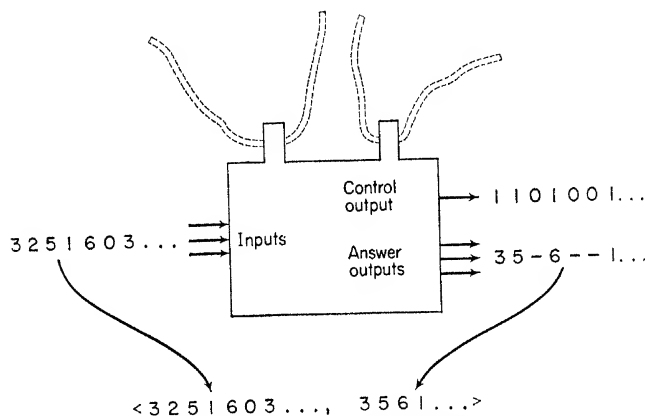


FIG. 2.2. Computation of generalized Turing machine.

number of ways to obtain what we will call a *generalized Turing machine*. To relate Turing machines to the fixed automata previously discussed we will allow the fixed part of the machine to have inputs and outputs, as in Fig. 2.2; as a consequence the earlier concept of *behavior* applies directly to Turing machines. Next we will allow

any finite number of tapes. Finally, we will regard a tape at any given moment of time as a finite automaton and construe a change in tape length as a change in this automaton; thus for us the tape is finite at each moment of time. A tape "square" storing one bit (a "0" or a "1") can be synthesized from switch and delay elements in a number of ways, and squares can be added or subtracted to represent the growing or shrinking of the tape.* Hence while a generalized Turing machine has a given structure at any moment of time its structure changes through time in a very simple fashion and under the control of the computation going on in the machine. A digital computer with indefinitely expandable magnetic tapes and a printed output is a generalized Turing machine.

Computation requires a problem and an answer, and the answer must be presented somewhere. Turing used alternate squares of the tape for this purpose. We will stipulate that the answer of a generalized Turing machine is to appear on the outputs. We will not, however, take the output history as the answer for to do this would be to identify behavior and computation, and hence to unduly restrict the domain of the computable. Instead, we will distinguish a *control output wire* from the remainder of the output wires, which will be called *answer output wires*. A "1" on the control output wire signifies that the state of the answer output wires at that time is part of the answer, while a "0" on the control output wire means that the state of the answer output wires at that time is to be ignored. The subsequence of the states of the answer output wires so selected is the *computed output sequence*. The sequence 3561 is the first part of the computed output sequence of Fig. 2.2. A pair consisting of an input history and the resultant computed output sequence is a *computation pair*. The *computation* of an automaton is the set of all computation pairs of that automaton.

There is a relational analogy between behavior and computation on the one hand and "real time" simulation and "non-real time" simulation on the other hand. In real time simulation the answer is produced at a rate determined by the actual time taken by the process which the computer is simulating, while in a "non-real

* See Section 4 below and also Section 6 of "The logic of fixed and growing automata," presented at the Harvard Symposium on the Theory of Switching, 1957, to be published in the *Proceedings* of the Symposium, edited by H. AIKEN.

time" simulation, the answer is produced at a rate independent of actual time. Similarly, the behavior of an automaton includes the outputs at every moment of time, while the computation includes only outputs selected by the computer itself, and in general these answer states do not appear at a uniform rate.

A computed output sequence may be null, finite, or infinite. A computation is said to be *infinite* if every computed output sequence (i.e. the second element of each computation pair) is infinite, *finite* if every computed output sequence is finite or null, and *mixed* otherwise.

We can apply the concept of computation (and the related concepts just defined) to fixed automata by distinguishing a control output of a fixed automaton from the other (or answer) outputs. Hence all three concepts—structure, behavior and computation—apply to both fixed automata and generalized Turing machines. These three concepts will also apply to the growing automata defined later (Section 4).

3. ANALYSIS AND SYNTHESIS OF AUTOMATA

Our discussions of growing automata in Sections 4 and 5 will presuppose some knowledge of fixed automata and generalized Turing machines, and so we will mention briefly some results about these special cases of growing automata and will note some of the unsolved problems. Since the concepts of behavior, structure and computation apply to growing automata, all the results and problems discussed in the present section apply with appropriate modifications to growing automata also.

Much is known about the behavior of fixed automata. For example, each fixed automaton has a finite number of delay output states, which fact has important consequences for the behavior of these devices. In the case of a fixed automaton without inputs it means that the behavior of the automaton is periodic.* In the case of an automaton with inputs the finitude of states means that the device can only detect "regular events,"⁽⁵⁾ that if the number of delay states is known then its behavior can be computed from the results of a finitely long behavioral test, and that there is a decision procedure for behavioral equivalence.^{(6)†} Since the notion of

* BURKS and WRIGHT, Ref. 3, Theorem I. Note that in this case the behavior is a single output history.

† BURKS and WANG, Ref. 4, Sec. 2.2.

computation is a new one for fixed automata little is known about this. The same argument that shows that the behavior of a fixed automaton without inputs is periodic shows that the computation of such an automaton is also periodic. There is an algorithm for deciding whether the computation of a fixed automaton (with inputs) is infinite, finite or mixed,* but the problem of the existence of an algorithm for deciding whether two fixed automata have the same computation is unsolved.

A great deal is known about Turing computability and the closely related notions of algorithm, general recursiveness and partial recursiveness. We do not have space to explain these concepts here† but it is worth noting that these results can be expressed in terms of our notion of a generalized Turing machine, and they contain answers to most questions about the computation of such machines. As applied to generalized Turing machines these results on recursiveness, algorithms, and computability need interpretation, however, because there are many ways of relating a generalized Turing machine to these notions. Thus there are alternative methods of using a generalized Turing machine to compute a recursive function. For example, for a given non-relative recursive function one can design a machine without inputs whose computation will be an enumeration of the values of that function. Alternatively, one can design a generalized Turing machine with inputs so that whenever arguments (problems) are presented on the input the computed functional values (answers) will appear as part of the computation. The distinction between finite or mixed computation on the one hand and infinite computation on the other (see Section 2) can be made to correspond to the distinction between partial recursiveness and general recursiveness.

We will relate a few of the things known about computability and recursiveness to our concept of a generalized Turing machine. Every computable number (in Turing's sense) can be produced as the computed output sequence of a generalized Turing machine with no inputs and one tape. The values of a function which is general recursive relative to a given function can be produced as the computation of a generalized Turing machine which receives the values

* Unpublished results of the author and Jesse Wright.

† See Kleene, Ref. 2 and Davis, Ref. 2.

of the given function as inputs. There is no algorithm which can decide of an arbitrary generalized Turing machine whether its computation will be finite, mixed or infinite, but there is an algorithm for deciding whether the structure of a machine can ever change. There is no algorithm for deciding whether two generalized Turing machines produce the same computation.*

Since the notion of behavior as applied to a Turing machine is a new one there are many open questions in this area. For example: is the class of behaviors of generalized Turing machines with $n + 1$ tapes larger than the class of behaviors for machines with n tapes? Is there an algorithm for deciding whether two generalized Turing machines have the same behavior?

Later we will discuss von Neumann's universal constructing machine, so Turing's universal computing machine (Fig. 3.1) is of

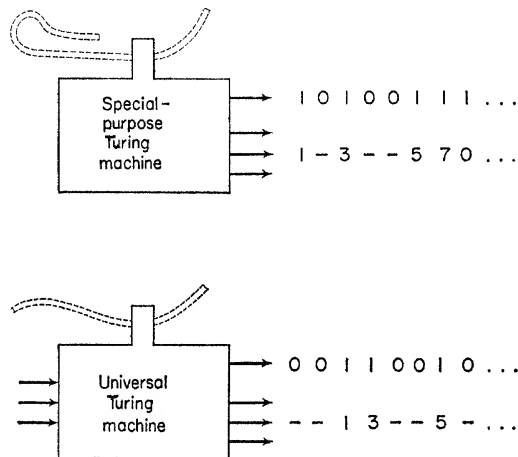


FIG. 3.1. Universal Turing machine.

particular interest here. Let us call a generalized Turing machine with no inputs, one tape, and one answer output wire a special purpose machine. Translated into our terms Turing's result is that there is a one tape generalized Turing machine with inputs, called a universal Turing machine, with this property: for each special purpose machine M there is a finitely long program or input sequence

* Unpublished result of Jesse Wright.

S such that when S is supplied to the inputs of the universal Turing machine the computed output sequence of M and the universal Turing machine are identical. This is illustrated in Fig. 3.1. It should be noted that in general the universal Turing machine and the special purpose machine do not produce the same output histories, so that the equivalence between them has to do with computation and not behavior. It should be noted also that the finitely long input sequence S can be supplied to the universal Turing machine by connecting a non-input cycle free sequence of delays to each of the inputs of the universal machine; the input to these is fed by a contradiction (an output that is always false) and the program is stored in the initial states of the delays. After a period of time equal to the length of this sequence of delays the input to the universal machine is always zero.

It is very likely that a similar result holds when inputs are added to the special purpose machine and more than one tape is allowed; that is, for each set of generalized Turing machines with n inputs and m outputs there is a universal Turing machine. Does an analogous result hold for fixed automata? Computer engineers often speak of trading time for hardware and of the choice between constructing ("wiring in") an operation and instructing (programming) that operation, and this might seem to imply the existence of a universal fixed automaton. But in fact there are no universal fixed automata. To prove this we first show that any fixed automaton A when supplied by a finite input sequence S produces a periodic output of period no greater than n , where n is the number of delay states of A . Since S is finite there will come a time when the input states to A are always the same. The automaton A then behaves as a non-input device with n delay states, and so the computed output sequence of A is periodic with period n or less. Hence A cannot produce a computed output sequence of, for example, period $2n$, and so A is not a universal machine. Thus there is no fixed automaton which is universal for the class of non-input fixed automata with m outputs, and consequently there is no fixed automaton for the class of fixed automata with n inputs and m outputs.

We will next mention some reduction problems. Does adding more tapes increase the class of computations produced by generalized Turing machines? A generalized Turing machine can have only one multi-channel reading and writing head per tape. In contrast,

the growing automata introduced in the next section can have tapes with more than one multi-channel read-write head attached to them. What effect, if any, does this have on the class of behaviors or the class of computations? In some cases one can use several single-headed tapes alternately to do the work of one many-headed tape. This can be done in the case of an automaton which records information on two tapes and compares these tapes, and perhaps this can always be done. It is worth noting that here we have an abstract buffering and scheduling problem.

No discussion of automata at a symposium on self-organizing systems would be complete without reference to machines which synthesize other machines. This synthesis process can be conceived in many ways. One method which has been studied some of late⁽⁷⁻¹⁰⁾ is illustrated in Fig. 3.2. Suppose we have a formal language L in

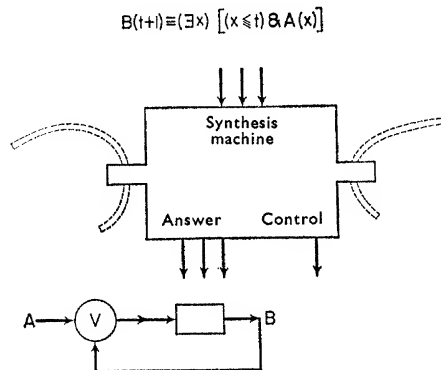


FIG. 3.2. Synthesis machine.

which a human can conveniently and precisely state some behavioral conditions C which he would like an automaton to satisfy. We would like to know if there is a synthesis machine S of which we can prove mathematically the following result: when S is given any behavioral condition C it will always produce either the design of some automaton A which satisfies this condition, or it will tell us that there is no such automaton.* The condition given at the top of Fig. 3.2 is that for all times other than zero the output of the

* Other problems in this area are formulated by Buchi, Elgot, and Wright, Ref. 7. What we have called a synthesis problem they call a combined solvability and synthesis problem.

desired automaton is to be active (in state 1) if the input has been active at some previous time. There are fixed automata satisfying this condition and so if a synthesis machine for a language containing this condition exists, this machine would produce as output a representation of one of these automata.

There is a large area of research here because there are many problems of the kind just described. The synthesis process represented in Fig. 3.2 starts with a particular language L and results in the synthesis of a fixed automaton satisfying a condition on its behavior. There are many different languages L which should be investigated and it is of interest to synthesize generalized Turing machines as well as fixed automata. In both cases we can allow the condition to be a condition on the computation of the desired automaton as well as a condition on its behavior. For each specification of the language L , the nature of the condition C , and the nature of the desired automaton A , the question arises: is there a synthesis machine and if so, what is its structure? In some cases there is a machine and in some cases there is not. It is important to realize that if there is no synthesis machine of this kind there is no method of using computers to design computers which is guaranteed to produce an answer in all cases. Hence the broad problem is this: to determine the theoretical limits of mechanizability of the synthesis process by deterministic machines of this kind.

4. GROWING AUTOMATA

John von Neumann has shown how to design universal constructing machines and self-reproducing automata within the framework of a general definition of automata.⁽¹¹⁻¹³⁾ The details of von Neumann's definition are contained in an unpublished manuscript, and so I have worked out an alternative definition. Von Neumann's model is much closer to biological phenomena than the definition presented here; for example, each of von Neumann's cells has a unit delay,* while many of our cells have no delays. Von Neumann's constructions are also based on very weak primitives (his cells are capable of only 29 different states),* while we have assumed very strong primitives (each cell may have any of about 28,000 structures and each structure is capable of several states). While the definition

* Shannon, Ref. 13, p. 127.

presented here is much less realistic and economical than von Neumann's, it does make it easy to construct and present various kinds of growing automata.

The basis of each growing automaton is an infinite two-dimensional array of discrete cells. At time 0 finitely many of these cells contain elements and the remainder are empty (structureless). The structure at time $t + 1$ is determined by the structure at t and the complete state of the device at t , according to rules to be explained. These rules do not allow more than a finite number of cells to be structured at any particular time.

There are four types of elements which may be created in a cell. Each element is binary in the sense that each wire or node of an element is capable of two states.

(1) A *computing element* may be a switch, a delay, a combined delay and switch, or just a plain wire. Each edge of a cell has at most one computing wire (shown in solid in the figures) impinging on it from within the cell. Specifically, the computing primitives are as follows; see Fig. 4.1. (a) Any switch with three inputs and

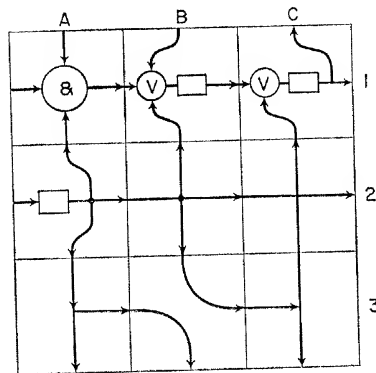


FIG. 4.1. Computing primitives.

one output, e.g. cell A1 of Fig. 4.1. There are 2^{10} of these, taking into account the different orientations of their outputs (north, south, east and west), and allowing all 256 possible three variable truth functions. Note that some of these truth functions will be independent of some of the inputs, so that this category includes as special cases two-input, one-input, and zero-input switches. This category

also includes certain wires, such as the wires of cells A1, B1, and C1 of Fig. 4.5. (b) There are unit delay elements with one output and one, two, or three inputs. These inputs come from one, two, or three edges through a disjunction. There are 28 of these; see cell B1 of Fig. 4.1. (c) There are unit delays with two outputs and one or two inputs (e.g. cell C1 of Fig. 4.1). There are 24 of these. (d) There are unit delays with one input and three outputs (A2 of Fig. 4.1); there are four of these since the input can come from one of four different cell edges. (e) There are wires running through a cell from edge to edge which involve more than two edges. Incoming wires may not merge and at most one wire contacts a side. There are 40 possibilities; see cells B2, C2, A3, B3, and C3 of Fig. 4.1. Altogether there are 1120 different computing elements.

(2) There are *scaffolding elements*, which have one input wire and either one or two output wires. Scaffolding wires are shown dashed; see cells B1, C1, B2, and C2 of Fig. 4.2. If we count primitives with different orientations as different primitives there may be any of 24 different scaffolding structures in a cell. The scaffolding primitives will conduct construction signals. A branched scaffolding element (e.g. cell B2 of Fig. 4.2) has an implicit switch

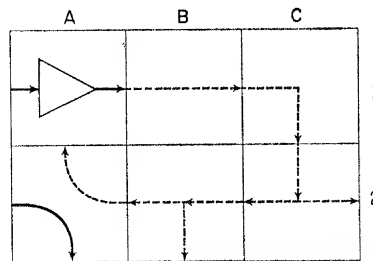


FIG. 4.2. Scaffolding and constructing primitives.

associated with it. By convention this switch has an initial setting; the switch setting may be changed by using switching sequences which will be introduced later.

(3) Any scaffolding element may coexist with any computing element in the same cell. See, for example, cell A5 of Fig. 4.4. There are 24×1120 such primitives.

(4) The last element is the key to the construction process. It is called a *constructing element* and is shown in cell A1 of Fig. 4.2.

A constructing element receives construction signals from a computing structure and routes these into a scaffolding structure; the construction signals travel through the scaffolding and eventually cause a cell to be structured in a way to be explained. A constructing element may receive its input from any cell edge and transmit its output to any other edge of that cell; hence there are 12 different constructing elements.

A cell may have an element of any of these four types or it may be structureless. Altogether there are 28,049 (which is less than 2^{15}) possibilities, each of which will be represented by a binary sequence called a *constructing sequence*. These sequences will be constructed or stored in a computer structure and fed into the constructing element. There is no upper limit to the length of computer wires or scaffold wires, and we will make the idealized assumption that signals will travel down the wires instantaneously.

The mode of operation of the system and the formulation of further rules can best be presented by means of an example. We will show how a tape may be lengthened under the control of a machine, thereby relating growing automata to generalized Turing machines. Figure 4.3 shows two "squares" of the tape, each

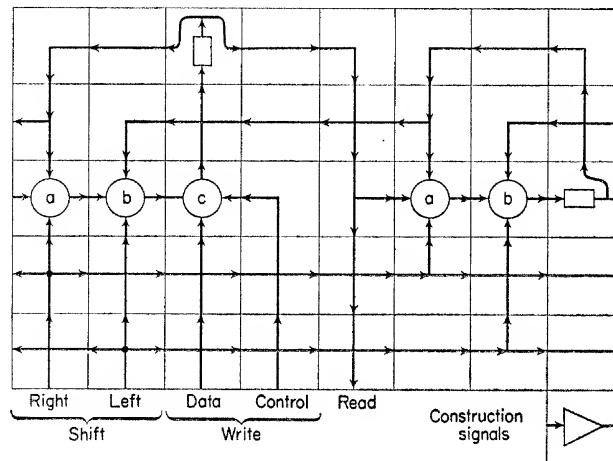


FIG. 4.3. Two "squares" of tape.

square storing a single bit. The switching functions (*a*), (*b*), and (*c*) are the following. Let "N", "S", "W", and "E" stand for

the north, south, west and east edges of a cell respectively. The output of switch (a) is E and is given by the equation

$$E(t) \equiv [\{S(t) \& N(t)\} \vee \{S(t) \& W(t)\}]$$

The output of switch (b) is E and is given by the equation

$$E(t) \equiv [\{S(t) \& W(t)\} \vee \{S(t) \& N(t)\}]$$

The output of switch (c) is N and is given by the equation

$$N(t) \equiv [\{E(t) \& W(t)\} \vee \{E(t) \& S(t)\}]$$

You are to imagine that a fixed computer which is associated with the tape is connected to the inputs and outputs of Fig. 4.3. This computer may "write" on the left-hand square and "read" what is stored on that square. It may cause the information on the tape to be shifted to the right by stimulating the "right shift" signal, and in a similar way it may cause the information to be shifted to the left. By feeding signals into the constructing element it can cause the creation, destruction, or alteration of new tape squares to the

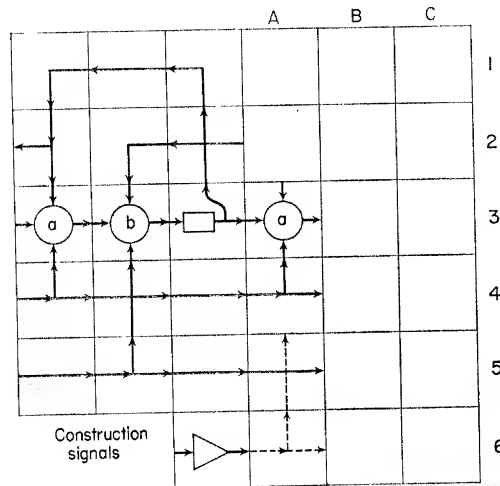


FIG. 4.4. Construction process.

right in a way to be indicated. A similar arrangement exists for modifying the tape to the left.

Let us next describe the construction of a square of tape. Cell A6 of Fig. 4.4 needs a scaffold with a branch. The construction sequence representing this element is produced by the fixed computer and fed into the constructing element, which produces the desired element in the cell touched by its output. The next constructing sequence will pass through the constructing element into the scaffold, through the vertical output of the scaffold, and will produce a new element in the cell (A5) which is touched by the output of the scaffold. In this case the structure produced is a horizontal computing wire and a vertical scaffolding wire. This process is repeated to give a structure to the next two cells (A4 and A3); the scaffolding of these cells is not shown because of pictorial difficulties.

After the rest of column A is structured construction activities need to be shifted to column B of Fig. 4.5. This rerouting is done

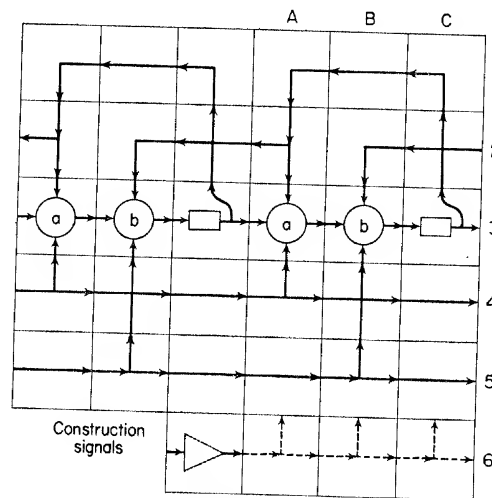


FIG. 4.5. Completion of new tape "square."

by a binary *switching sequence* which goes through the constructing element into the scaffold and changes the switch setting of the scaffold of cell A6 so that subsequent constructing sequences are routed into cell B6. After this we can structure column B by means of six constructing sequences. Another switching sequence resets the switches in cells A6 and A5 and permits the structuring of column C.

This switching sequence works as follows. It would contain two switching signals in coded form. The first signal would be received by A6, and in this case would leave the switch set as it was. The second switch signal would be passed on to B6, and would cause its switch setting to be changed so that the next constructing sequence would effect cell C6 (rather than travelling through the B column). In general a switching sequence will contain as many switching signals as there are switches which must be passed through to get to the last switch to be modified.

The central computer may want to contract the tape as well as to expand it, and so it needs the capacity to wipe out the structure of a cell. This is provided for by binary *destroying sequences*. A destroying sequence passes through the constructing element and down the scaffolding in the same way as a constructing sequence. However, a destroying sequence affects the cell which contains the terminus of this scaffolding channel rather than the next cell; it

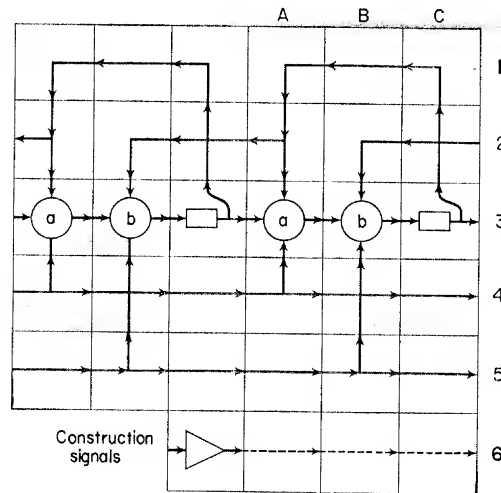


FIG. 4.6. Preparation for next construction.

causes the element of this cell to be destroyed. There is a sequence of 18 destroying sequences and 2 switching sequences which will destroy the tape square just created.

We may also use a switching sequence followed by three destroying

sequences to erase the elements of the bottom row (6) and then use 3 constructing sequences to create the scaffolding shown in Fig. 4.6. Thus with a master sequence of 21 constructing sequences, 3 switching sequences, and 3 destroying sequences we end up with a new tape square and with the scaffold arranged so that each repetition of this master sequence will produce a new square to the right. By iterating this master sequence we can produce a tape of any length for a generalized Turing machine. The process of construction just described takes many moments of time to produce one tape square, and hence is not as fast as that stipulated for the generalized Turing machine, which can take place at the rate of one square per moment of time. We could handle this discrepancy in either of two ways. First, we could use the growing automata to simulate a generalized Turing machine. To do this we associate one major cycle (i.e. the time required for the construction of a tape square) of the growing automaton just described with one minor cycle (i.e. a single unit of time) of the generalized Turing machine. Second, we could enrich our base of construction by adding elements and wires capable of more than two states. It would not be necessary to introduce computing elements with more than two states. It would suffice, for example, to let the scaffolding elements have many states and to have constructing elements which convert binary signals into signals properly coded for the scaffolding elements; note that this method would not involve any increase in the number of computing primitives or the number of scaffolding primitives, and not a very great increase in the number of constructing primitives.

We have said enough about the definition of growing automata to make its main features clear. Various details need to be specified before the definition is complete. We will assume that a primitive element is active as soon as it is created. This means that part of a computer which has been constructed may start to operate before the computer is finished. This fact needs to be taken account of in the design of the computer, as it often does in the design of actual computers, since these computers may have to be cleared to a standard initial state before they are ready to operate. Two constructing sequences, coming from two different constructing elements, might arrive at a cell at the same time; a priority rule is needed to determine which one governs the construction. There are

other minor details which need to be specified but we will not bother with these here.*

Let us consider next what automata can be specified as growing automata in the sense of the present definition. Every fixed automaton can be put into a normal form,[†] and this normal form can be constructed of computing elements in cells. So each fixed automaton is an unchanging growing automaton. We have shown in some detail how an expanding and contracting tape may be constructed.

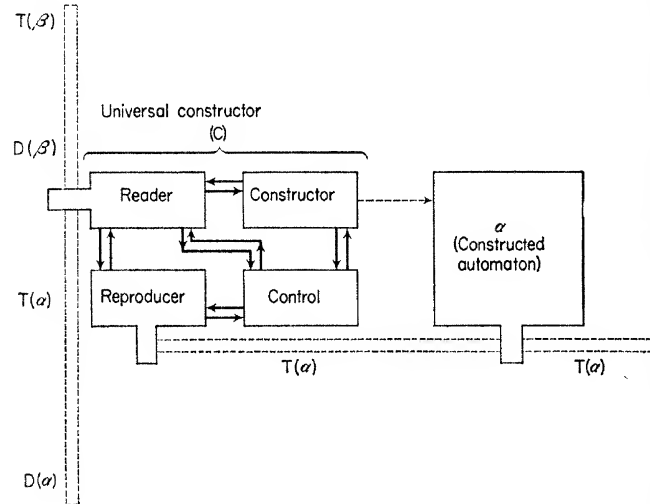


FIG. 4.7. Universal constructor.

Hence, except for a matter of speed, every generalized Turing machine is a growing automaton.

One of the growing automata which can be constructed is von Neumann's universal constructor, of which a block diagram is

* It is perhaps worth noting that the definition does not prevent the construction of automata which are not well-formed (e.g. have switch cycles), nor is there an effective criterion for deciding whether a given growing automaton will ever "develop" switch cycles. Restrictions could be added to guarantee that every automaton is well-formed. The simplest such restriction would be to incorporate a delay in every switch, but this tremendously complicates the automata and limits their behavior. A less severe restriction is to require every switch or wire with an output on the south edge to contain a delay. We will not take time here to investigate the various possibilities.

[†] Burks and Wang, Ref. 4, Section 2.3.

shown in Fig. 4.7. The *universal constructor* C is a finite structure to which may be attached a tape containing the specifications of any number of automata which are to be synthesized. In general, each specification consists of two parts. The first part of the specification is a master sequence of constructing sequences, switching sequences, and destroying sequences, which will cause a fixed part of the desired automata to be constructed. This master sequence is the description $D(\alpha)$. The second part of the specification contains the contents $T(\alpha)$ of a tape which is to be constructed and attached to automaton α .

The universal constructor operates under the direction of its control as follows. First, the tape reader reads $D(\alpha)$ and transmits it to the constructor; concurrently the constructor constructs α . Next, the reader reads $T(\alpha)$; concurrently the reproducer produces a tape, attaches it to α , and records $T(\alpha)$ on it. This process is then repeated, so that at the end of the construction process $T(\alpha)$ is recorded twice on the tape. (The reason for this duplication of tape information will become apparent in a moment.) In general, for each description $D(\alpha)$ of an automaton and each tape content $T(\alpha)$, the

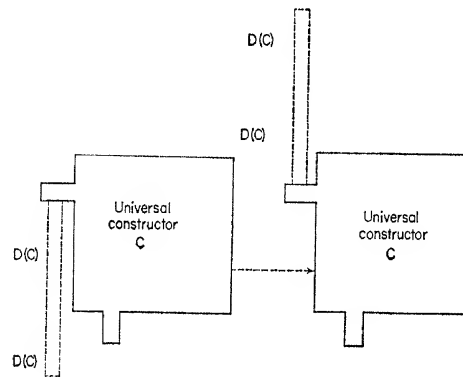


FIG. 4.8. Self-reproduction.

universal constructor will construct α and record $T(\alpha)$, $T(\alpha)$ on the tape attached to α .

As von Neumann showed, the universal constructor C together with the proper tape will reproduce itself; see Fig. 4.8. Let $D(C)$ be a sequence of constructing sequences, switching sequences, and

destroying sequences which completely describe C (including its scaffolding). Let the universal constructor C contain a tape storing $D(C)$ twice when it begins action. The universal constructor will first read $D(C)$ and produce C . It will then read $D(C)$ twice and produce a tape recording $D(C)$ twice. (This is the reason for the earlier requirement that $T(\alpha)$ be copied twice.) Thus the universal constructor C when supplied with a tape containing $D(C)$, $D(C)$ produces a universal constructor plus a tape containing $D(C)$, $D(C)$. Hence C with a tape storing $D(C)$, $D(C)$ reproduces itself. The new combination of C and the tape $D(C)$, $D(C)$ can then reproduce itself, and this process can be repeated *ad infinitum*.

If we combine a universal constructor and a universal computer (i.e. a universal generalized Turing machine), and supply this composite machine with a tape containing its description twice, this

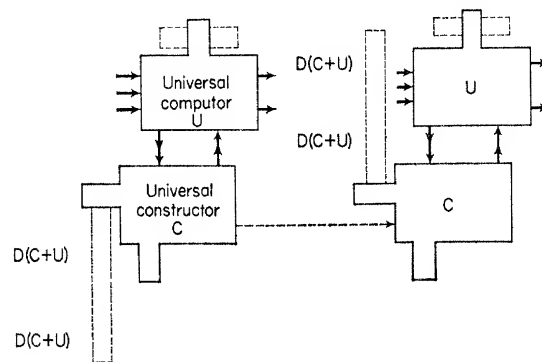


Fig. 4.9. Universal constructor-and-computer.

whole automaton will reproduce itself; see Fig. 4.9. Thus a universal constructor-computer can reproduce itself.

5. SOME PROBLEMS ABOUT GROWING AUTOMATA

Since growing automata include fixed automata and generalized Turing machines as special cases, the problems described in Section 3 apply to growing automata also. There is another type of problem, quite different from any of the problems discussed before, which is applicable to automata of all kinds; this concerns the relation of the cyclic complexity of the structure of an automaton to its behavior. Since our earlier discussion of a generalized Turing

machine did not make explicit whatever cycles are implicit in the tapes, this problem had to be postponed to the present section.

Let us consider the problem first for the case of a fixed automaton. We can divide the structure of a fixed automaton into *maximal cycles*. The *degree* of a maximal cycle is the number of unit delay elements in it, and the degree of a fixed automaton is the maximum of the degrees of its maximal cycles.* Figure 5.1 consists of a single

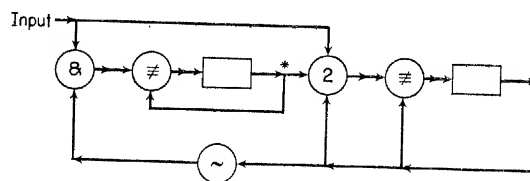


FIG. 5.1. Automaton of degree 2. Ternary counter.

maximal cycle of degree 2; the output nodes are marked by stars. Figure 5.2 contains six maximal cycles, each of degree 1, so Fig. 5.2 is of degree 1; its output nodes are marked by stars.

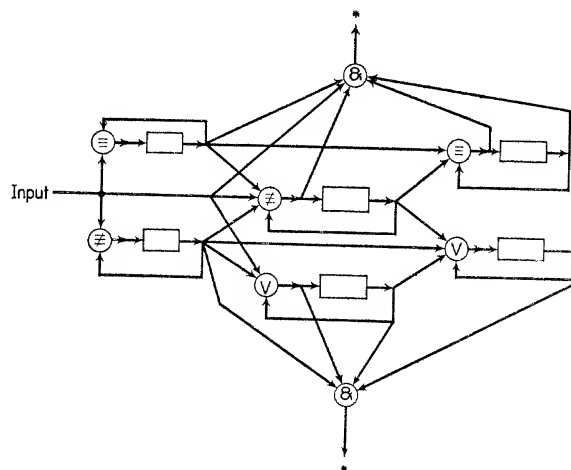


FIG. 5.2. Automaton of degree 1.

How, now, is the degree of an automaton related to its behavior? Is there some critical degree n , such that all fixed automata behaviors

* These concepts are explained in detail in Sec. 4.1 of Burks and Wang, *op. cit.*

can be realized by structures of that degree or less? It has been conjectured that there is no such degree. The general problem remains open, but John Holland has shown that no automaton of degree 1 can realize the behavior of Fig. 5.1, which is a ternary counter.⁽¹⁴⁾ ✓

The same problem can be raised about the computation of a fixed automaton: is there some critical degree m , such that all fixed automata computations can be realized by structures of that degree or less? Both problems can be extended to generalized Turing machines and to growing automata generally. Of course the structure of any growing automaton changes with time, and hence the degree of the structure changes with time as well. But we can define the degree of a growing automaton to be the maximum degree of any of its structures at any time for any input history, if the maximum exists; otherwise the degree is infinite.

We can now ask these questions about growing automata: (1) Is there some finite degree n such that all behaviors can be realized by growing automata of this degree or less? (2) Is there some finite degree m such that all computations can be realized by growing automata of this degree or less? It might seem that the answer to both these questions is "no," since clearly in both cases there is no upper limit to the length of the tapes needed, and the tape designed in the preceding section is of infinite degree. However, it is possible to design a tape of finite degree by using isolated squares and moving the read-write head along the tape instead of shifting the information along the tape. To store a 1 in a square the automaton builds a delay cycle in which a pulse circulates; to store a 0 it builds a delay cycle whose successive states are 0. The fixed part of the machine can send out a long wire to do this and by means of another wire can sense the contents of the delay cycle. The contents of the square can be changed by destroying the delay cycle in it and rebuilding the appropriate one. When the fixed part of the automaton is neither reading from the square nor writing in the square it would have no connection to it. In this way a generalized Turing machine may be designed which has a fixed degree.

There are other measures of feedback complexity which are of interest. Consider a fixed automaton with n unit delays $x_1, x_2, x_3, \dots, x_n$. Let $f(x_i)$ be the number of delays which directly and immediately affect the delay element x_i , i.e. whose outputs drive

the input of x_i through switch elements. In Fig. 5.1 $f(x_1) = f(x_2) = 2$ since each delay drives itself and the other delay through a switch. Define an average measure of feedback complexity F for the automaton by

$$F = \frac{\sum_{i=1}^n f(x_i)}{n^2}$$

There are "directly entirely connected" automata for which $F = 1$,* Fig. 5.1 being an example. However, the amount of feedback complexity F in actual systems is usually much less than the maximum. $F = 7/18$ for Fig. 5.2. F is small in a typical modern digital computer. (Note that such a computer is pretty nearly one maximal cycle, not counting the tapes.) This fact is relevant to our ability to understand a large digital computer. Because the ratio of actual feedback lines (from delays) to possible feedback lines is low the various parts of the computer have a large amount of local autonomy and may be understood in isolation from the rest of the computer. (Uniformities within these parts also make it easier for us to understand them.) It is worth noting that if neurons are represented by switches and delays, F is small for the human neural system.

Let us turn next to some problems which are unique to growing automata. The first is one that von Neumann considered: How complex does a self-reproducing automaton need to be? In his model von Neumann used cells which were capable of about thirty states, and he estimated that about 200,000 cells would suffice for self-reproduction. The cells of our models are much more complicated, for a cell may contain any of about 2^{15} different elements, and each element is capable of several states. But counting states is not sufficient, because there are many ways in which states may be used. For example, a primitive element capable of transferring its structure to a neighbouring cell on receipt of a single pulse is much more powerful than any of the primitives we have assumed.

One could compare different models of growing automata by

* See BURKS and WANG, Ref. 4, p. 291. Of course every automaton can be put in a normal form which has this property, but we are here interested in nets in "reduced form" in which all irrelevant (noneffective) switch input wires are deleted.

reducing them to a common system. For example, we could simulate both von Neumann's model and ours by means of infinite homogeneous fixed nets of switch and delay elements. We will indicate how this could be done for our definition. Each cell would be simulated by a fixed net inside a square; these squares would be repeated *ad infinitum* in the same way that the cells of our model are. Each fixed automaton in a square would contain a register, a master switch, and switches and delays which would perform the scaffolding and computer functions of the original cell. Each fixed net of a square would have inputs to and outputs from the four neighbouring squares; these inputs and outputs would be used for scaffolding wires and computer wires. A constructing sequence which was to affect a given square would be routed to the register of that square; the register would then determine the connections of the switches and delays by means of the master switch; in this way the fixed automaton of the square would simulate the original cell. Flip-flops would be used to store the switch settings of branched scaffolding elements; a switching sequence passing through a square would set the flip-flop correctly and then pass on to set the flip-flops representing other scaffolding branches. A destroying sequence would go into the register and restore the square to its original state.

Note that only one infinite homogeneous structure (namely, an infinite repetition of the fixed automaton in a square) is required for our definition of growing automata. Different automata are represented (simulated) by different states of this structure, so we are here using automata behavior to simulate automata structure.

Von Neumann's model could be simulated in a similar way. Then the complexity of a von Neumann self-reproducing automaton could be compared with the complexity of the self-reproducing automaton we described by making a weighted count of the number of switches and delays used in the initial structure of each. By precisely defining a common reference system in this way we can make the problem of minimal complexity of cell reproduction into a strictly logical problem. Experience with minimality problems suggests that it would be extremely difficult if not impossible to find the minimal solution and prove it minimal. One might, however, obtain an estimate of minimal complexity by using a computer to help work out many alternative constructions.

There are many other concepts of growing automata worthy of investigation, and they will give rise to new problems. In his first model von Neumann had girders, sensing organs, and joining organs, as well as computing elements, floating freely in a lake. As Shannon remarks,* because of the complexity of motion which is thereby possible it would be exceedingly difficult to give a detailed design of a universal constructor in this model. One might, however, augment the definition of growing automata we have given to obtain some of the effects of von Neumann's first model without all the complexity. For example, we might add a *sensing element* which would sense the structure and contents (state) of a cell. The sensing element would send *sensing channels* through the cells in a way similar to the way a scaffold is constructed by the constructing element. These sensing channels would not disturb the structure already in the cells, but would carry information about the cell being sensed back to the sensing element. By a process analogous to the construction procedure described in the previous section, an automaton could sense what element was contained in a cell and the state of that element at that time. This information could be used in various ways, e.g. it might be used to detect a fault which could later be repaired by the constructing element.

The question then arises: How much self-knowledge can an automaton acquire and keep? This question is of interest to the theory of self-organizing systems since a possible procedure in self-organization is for the system to have a picture of its structure and behavior and use this picture to guide itself. Clearly, there are limits to how much knowledge an automaton can acquire and keep about itself. For example, a growing automaton might change faster than its sensing elements could sense these changes. There are also logical limitations. Thus if all the states of an automaton are used, a proper part of the system cannot have a completely detailed picture of the whole, for the whole, being larger than the proper part, would be capable of more states than this part.

The logical paradoxes of self-reference are relevant here. It is well known that some referential statements lead to contradictions.

* Ref. 13, pp. 126-127.

Thus the sentence, "This sentence is false," is false if true (because it says of itself that it is false) and true if false (because if it's false that it's false then it is true), and hence both true and false. To resolve such paradoxes Russell proposed⁽¹⁵⁾ a formal language in which no self-referential statements could be formulated. The proposal that no self-referential statements be allowed in a formal system is much too strong, however, for there is clearly nothing wrong with the self-referential statements: (1) "This sentence is in English" (2) "This sentence is written at least once." Indeed, both of these sentences are true. These sentences illustrate the fact that a sentence can refer to its physical properties without contradiction. There are also cases where a sentence can refer to its structural or syntactical properties without contradiction. The sentence, "This sentence is grammatical," is a case in point. Another case of a formula consistently referring to its own structure is the undecidable formula constructed by Gödel* in his proof of the incompleteness of arithmetic. This is a formula F which says, in a certain sense, " F is not provable."⁽¹⁶⁾ Provability is a purely syntactical concept so this formula is referring to its own syntax, and it does so without contradiction.

Thus a sentence may contain a description of some of its own syntactical or structural properties. Note the resemblance of this to what we found in the preceding section, namely, that a self-reproducing automaton C contains within itself a description $D(C)$ of its own structure. It does not, however, contain a description of its own state of contents (except for its homogeneous initial state). This suggests that an automaton with a sensing element might be able to sense its own structure and construct and store a description of its structure. There is a problem in designing an automaton to do this, however.

Prima facie it might seem that an automaton could not store a description of its own structure because, however many cells it had, storage of the description would require more than that number of

* In his lectures at the Institute of Advanced Study in 1933-34 (as reported in some mimeographed notes) Gödel pointed out that his undecidable formula showed that Russell's solution of the paradoxes was too extreme because the undecidable formula (truly) says of itself that it is not provable. He agreed with Russell that there must be some limitation on what a sentence may say about itself and suggested that no sentence can properly talk about its own truth and falsity.

cells, in the manner of the Tristram Shandy paradox.* This objection is of course not sound, because we may use indices, summation signs, and quantifiers in the description. You will recall that a certain master sequence is used to describe one square of the tape. In writing the description $D(\alpha)$ of the structure of automaton α one does not repeat this master sequence for each square of tape to be constructed, but rather uses the same sequence repeatedly in a manner familiar to all programmers. This can be done because the tape has a highly uniform structure.

It follows from these considerations that in order to produce a description of its own structure which is sufficiently small to be stored in that structure an automaton must detect some of the uniformities of its structure. It could, for example, systematically sense the structure of each of its cells and then analyze the results in order to find a more compact statement of them. Hence the amount of information a growing automaton can acquire and store about its own structure depends on what mechanical techniques there are for discovering uniformities of structure.

REFERENCES

1. J. von NEUMANN, Probabilistic logics and the synthesis of reliable organisms from unreliable components, in *Automata Studies*, pp. 43-98 (ed. by C. E. SHANNON and J. MCCARTHY), Princeton, 1956.
2. See for example, S. C. KLEENE, *Introduction to Metamathematics*, New York, 1952, and MARTIN DAVIS, *Computability and Unsolvability*, New York, 1958.
3. A. W. BURKS and JESSE WRIGHT, Theory of logical nets, *Proc. Inst. Rad. Engrs.*, **41**, 1357-1365 (1953).
4. A. W. BURK and HAO WANG, The logic of automata, *J. Assoc. Computing Mach.* **4**, 193-218, 279-297 (1947).
5. I. M. COPI, C. C. ELGOT and J. B. WRIGHT, Realization of events by logical nets, *J. Assoc. Computing Mach.* **5**, 181-196 (1958).
6. E. F. MOORE, Gedanken—experiments on sequential machines, pp. 129-153 of *Automata Studies*, *op. cit.*
7. J. R. BUCHI, C. C. ELGOT and J. B. WRIGHT, The nonexistence of certain algorithms of finite automata theory, Abstract, *Not. Amer. Math. Soc.* **5**, 98 (February, 1958).
8. A. CHURCH, *Application of Recursive Arithmetic in the Theory of Computers and Automata*, Notes from summer session course in Advanced Theory of the Logical Design of Digital Computers, University of Michigan (1958).

* Tristram Shandy took two years to write the story of the first two days of his life! (Laurence Sterne, *The Life and Opinions of Tristram Shandy, Gentleman*, vol. 4, ch. 13.)

9. C. C. ELGOT and J. R. BUCHI, Decision problems of weak second order arithmetics and finite automata, part I, Abstract, *Not. Amer. Math. Soc.* **5**, No. 7, 834 (December, 1958).
10. C. C. ELGOT, Decision problems of weak second order arithmetics and finite automata, part II, Abstract, *Not. Amer. Math. Soc.* **6**, No. 1, 48 (February, 1959).
11. J. von NEUMANN, The general and logical theory of automata, in *Cerebral Mechanisms in Behavior—The Hixon Symposium*, pp. 1-31 (ed. by L. A. JEFFRESS), New York, 1951. This is reprinted in *The World of Mathematics*, vol. 4, pp. 2070-2098 (ed. by J. R. NEWMAN).
12. J. G. KEMENY, Man viewed as a machine, *Sci. Amer.* **192**, 58-67 (April, 1955).
13. C. E. SHANNON, von Neumann's contributions to automata theory, *Bull. Amer. Math. Soc.* **64**, 123-129 (May, 1958).
14. J. H. HOLLAND, Cycles in logical nets, University of Michigan Ph.D. dissertation, 1959. This thesis contains many other results in the same general area.
15. B. RUSSELL and A. N. WHITEHEAD, *Principia Mathematica*, Ch. II of the Introduction ("The Theory of Logical Types").
16. KURT GÖDEL, Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, *Monat. Math. u. Phys.* **38**, 173-198 (1931).

DISCUSSION

MINSKY: I wonder if your last statement is associated with any kind of proof? It seems to me it is quite possible for an automaton of a rather simple kind to find out all about its structure, what its structure was at a certain moment, simply by making a sort of rubber stamped copy of it, and so finding out its whole state at the same time.

Now, this is non-trivial, but it is a way of having each cell essentially have dual sets of states and there is a way which results from what you may have heard of as the soldier problem of sending waves which propagate through an automaton of one or two dimensional structure and will simultaneously transfer the entire structure of that automaton to another layer. At that point the machine can leisurely investigate every detail of what its structure was at a certain time after it gave the order to make a copy and this amounts in a sense to the idea of machine which not only has the kind of structure you described, but has a hand with an eye on the end of it and perhaps even has a mirror so it can even find out what the structure of the eye is. But furthermore, it can do this with a frozen copy of itself so it can see all the states. So I really have a serious question in my own mind and I wonder if other people do, if you can make any such generalizations connected with self-reproduction that says, to understand yourself or know your state, you really have to make generalizations. I couldn't see how von Neumann showed this and I still don't think there is much evidence for it.

BURKS: I was talking about the kind of situation where one designed a general purpose computer of not very uniform structure (though there would obviously be some uniformities) which had, for example, a single sensing element. The problem is to write a program so this sensing element will make the proper survey and the automaton will end up with a description of itself stored in itself. I didn't mean to imply that this couldn't be done, but I did want to say that in this general case it is necessary that the automaton have some ability to summarize its description.

SOLOMONOFF (*Zator Company, Cambridge, Massachusetts*): I have some question about the motivation for studying self-reproducing machines, and growing automata. It would appear to me as far as the practical problem of getting self-reproducing machines or factories is concerned, that we can from the theoretical point of view do this fairly well without that type of consideration. It would appear from a practical standpoint that we could construct self-reproducing machines or factories without making a consideration such as the kind you have made and that the problems are essentially engineering problems.

From another point of view, the fact that a man is self-reproducing tells us very little about its structure, I think, and so, from this standpoint, this isn't a particularly good way to get insight on the way living things in general work, so I just am wondering what sort of motivation you have for studying this other than a mathematical one?

BURKS: Well, I think the last thing you said is sufficient: the mathematical motive. I can agree for purposes of this response that this mathematical approach might not be a useful way of studying natural methods of self-reproduction. Certainly the particular model I propose is highly idealized in this respect and even the ones that von Neumann considered were somewhat idealized. But I think the general processes are of interest here and in particular I tried to relate the operation, structure, computation, and behavior of self-duplication, automata to other automata which are widely studied: Turing machines, which grow in a very simple manner (the tape growing in a uniform way) and fixed automata. The objective of course, would be a comprehensive theory of automata which would include all these as special cases.

KIRSCH (*Bureau of Standards*): You gave an example in your talk of a quantification scheme which can be realized as you showed by a particular logical net and as you implied or pointed out directly, clearly there are such quantification schemes for which there is no realization. I think if your particular example, if you reversed the inequality—you have a statement that a form exists, such that X is greater than T . That is the predictor for which you obviously couldn't synthesize an organism.

Now my question is this, can you say something about the extent to which you can cover a wide—the extent to which you can say of a fairly wide class of quantification schemes—whether or not they are in fact representable by a finite automaton? What I am looking for is a statement of this sort for lots and lots of cases, almost all cases. Or all the important cases we can do this, of the converse statement, or neither or both.

BURKS: As you remarked, it is very easy to write down formulas that can't be realized because they are asking the automaton to predict the future and this is not within its powers. It has also been proved that not all automata can be described by quantifier formulas of this sort. (I say "of this sort" because at the time at my disposal I didn't have time to give a formal definition of a language which would allow this and other reasonable expressions.)

You might consider some language which is strong enough to include bounded quantifier formulas but is not much stronger. It is unknown for any such language whether there is a synthesis machine that will tell you if there is an automaton satisfying a given condition and if there is such an automaton will describe it to you.

I don't think it is a very serious objection that there are lots of formulas to which no automata correspond. If the designer writes down something which states a condition that no automata can realize, then the machine has done its job if it says there is no such automaton.

There is a further complexity here in that you can increase the scope of the problem to include that of getting an automaton which computes "such-and-such" a thing, and from many formulas which anticipate the future you can construct automata which in a suitable sense will compute that sort of thing.

SECOND DAY

GENERAL DISCUSSION

Questions directed to the Panel at Large

KATZ (RCA): *Quo vadis?* What significant milestones lie ahead? Also speculate when they will be realized.

CAMPBELL: As an outsider I will offer a comment. It seemed to me that in the topics discussed by Goldman yesterday, where you have very rigid physiological systems, homeostatic systems in a very rigid sense, and a well-developed linear servomechanism theory, there is a juxtaposition that may pay off right within our own lives. I have a similar feeling that there is great promise in an engineering approach to locomotor automata coupled with a careful inspection of the evolutionary record. There is room for talented people like yourselves to look in detail at the evolutionary development of the nervous system from an engineering framework, asking in how many ways could one develop a system which would steer itself, what sources of information were available in the physical environment to the animal or machine at what particular evolutionary level, etc. Von Uexkuhl's old notion of the *Umweltlehre* is relevant. In spite of his mystical vitalism, he nonetheless had a very important insight, which I think a servosystem engineer could make concrete sense of: What primitive assumptions about the environment would machines have to have at each level to duplicate the evolutionary sequence?

BURKS: I am not a seer, so I won't attempt to foretell the future, but I would like to state one thing I would like the future to produce. The need for this, I think, has been illustrated in our symposium the last two days. An example is the question that was recently directed to Al Newell: "How much did you have to tell the general purpose learning machine about the problem before it could go on and do something with it?" Also, we have been asked: "What is a self-organizing system?" and "Isn't it really cheating when you tell it this much?"

The kind of thing that is of interest here is the fact that to answer these questions in the present state of knowledge one has to recite in detail what one has done. This shows a lack of general concepts that are relevant here, that could help us communicate in a group of this sort and say, much more briefly, how much we had to tell the machine, how much we are cheating, and to what degree the machine is self-organizing.

And so I feel a need for a theory and concepts of the general sort I have described that would help in the future; but I won't attempt to foretell whether it will appear in the future.

McCULLOCH: It seems to me that one of the most interesting questions and the one that keeps being asked is this. Devices working on random networks, picking up information, becoming organized in terms of contingent probabilities and so forth, rewarded by more material or time or components, more energy, whatever you will, can obviously learn to muddle through in the world. That is perfectly clear.

On the other hand, there is a peculiarity of our guesses that we talk in terms of noiseless concepts. By a circle we mean something whose ratio of circumference to diameter can be pushed to any number of decimal places. This sort of thing does not seem to me to be likely ever to arise so long as we are dealing with these devices as proposed at the present moment until one thing happens. Let us suppose that we have two machines with adequate memories and good ability at calculating contingent probabilities, that we teach each of these chess, without teaching them the rules, that we merely insist that they learn the rules by learning to play against an opponent.

The one thing that we build into each machine is a desire to play. It doesn't know what playing is except it starts moving pieces and it wants the other fellow to move pieces. That is to be its game. That is all it knows. Sooner or later it will learn the rules of chess.

Suppose we have two such machines that have arrived at this stage and one of them has to do some other chore, so they agree to take another machine and teach it the rules of chess. The minute this situation arises, in which they have to lay down the rules for playing the game of chess, they will create the noiseless concepts which are the rules for the game of chess. That is, it seems to me, that these noiseless concepts are not the consequence of thinking in natural terms, but of thinking in conventional terms and that until machines begin to play ball with each other to the extent of creating a language by which they can speak to each other. (I don't mean one we put into them, but one invented by them in which they can formalize and convey their own experience.) I don't think that until that time we have any right to look inside the machine for the equivalent of our noiseless concepts.

PASK: I can see the future of this field as lying with those machines that just can't help learning. You can name a hundred systems that given stimuli just can't help learning about their environment and playing a game with it. It is also easy to build such systems. Anybody can.

Now given such a system as this, we then want to know how can we interact with it, how can we make sense of it? That is the first question.

The second question, which I think is much more important, relates to playing a game with the system (or watching a game between sub-systems) which generates as Dr. McCulloch says, a noiseless concept.

The first question I believe to be immediately answerable, possibly using the technique, which I sketched over in my proposal to-day, for viewing the system as a natural historian. In the simplest case our scrutiny determines the functionally distinct sub-systems (usually different ones at different instants) into which the assembled may be partitioned. Thus the crummidge which learns and adapts and just can't help doing so, can be divided for descriptive purposes into component sub-systems which can be dealt with. In particular, the activity of some of these can be suitably "rewarded"—recalling that in this context I mean, by "rewarded," given permission to grow and use many of the raw materials.

Now having done this, we can apply these beautiful concepts of logical stability and ultra-stability to our model of the system. If this allows us to work out rewarding strategies we can afford to have confidence in the machine and I think in the process of the calculation. As to the time it will take—we are at a stage where—from a wholly practical point of view—we need a Calculus of Self-Organizing Systems. It would be a sad state of affairs if we don't have it in five years.

BURKS: I would like to make one more comment on this if I may. The idea of the noiseless concept is essentially the idea of a deterministic rule and you

X*

ROUNT LIBRARY
CARNEGIE-MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213

might put Dr. McCulloch's suggestion in another way: the machines will have to become aware of the use of deterministic rules.

TAUBE (*Documentation, Inc.*): Is similarity always reduceable to identity and difference? If the answer is yes, please give evidence for such an assumption. If similarity is a primitive notion, can a machine recognize similarity—that is, without reducing it to identity and difference?

NEWELL: I am not sure I can meet the conditions of the question about giving an example. The way it goes in all of our work, the answer is yes. In the basic programming language we have used an identity test on symbols. That is the only place you can ultimately make a differentiation. Consequently, at some level or other you write programs to do this. I guess it never occurs to us to ask the question whether all similarity can be reduced to this. We pick up little problems and we find good definitions of differences and good definitions of similarity that help us to solve the problems and the definitions always turn out to be things like this. So, as far as I am concerned, this is a kind of residual issue on which I take no stand in the long run, but just keep chipping away at it. I would say that fundamentally in our languages similarity is identity between a pair of symbols.

BURKS: It seems to me this is largely a question of definition. Take two colors that are similar, but not exactly identical. Of course they are different in that they are of different wave-lengths, but they are identical in the sense that they both belong to some class of wave-lengths that constitutes generic color.

CHAIRMAN MCCARTHY: Are the rest of the speakers satisfied with these answers?

MCCULLOCH: No, I am not satisfied with any answer to this question. The greatest puzzle I know anywhere in the world, is the puzzle which as a psychologist I am familiar with under the name of stimulus equivalence. The hardest thing I know is to find out what is it that a beast regards as the same in the sense that he will make the same responses to it. You may think it is one thing and the beast may be working with you quite happily from month to month and you suddenly discover there is an entirely different thing which to him is the same.

One of my associates removed a visual cortex from a monkey and it took him a long time to find out that thereafter the monkey was measuring the total luminous flux, a thing to which no monkey would ordinarily make a discriminate response. So, I am not at all happy. I suppose by "similarity" one means in this case what the psychologist means when he speaks of "stimulus equivalence."

Certainly in the statement " A is B and B is C , therefore, A is C ," there are two " A 's" on the page, and they are not identical. One is on one line and one is on another. Certainly with respect to their meanings they are supposed to be equivalent or one is in bad trouble with his logic; that is to say, they must be "stimulus equivalent." I suppose that is what we are talking about. I may be wrong.

BURKS: Doesn't this show that similarity is relative? A is similar to B , relative to some respect to C .

NEWELL: If you try to put this in a program, if you try to write out a program for this in the kind of machine we deal with, you construct something that is in effect an identity test. If you don't like to you needn't call it an identity test—but the test says: if your accumulator is zero, branch (and you make the accumulator zero by subtracting two things). Thus, in the way our language is written, the test always says: if two symbols are identical then branch in one

direction; if they aren't, take the other branch. So ultimately any such decision is made by abstracting until you get some symbols that should be identical.

MILNER: As a psychologist may I say a few words about this stimulus equivalence problem? As Lashley pointed out, there are two possibilities here, and they both operate. One of them, of course, is that two things may be similar because the stimuli fire overlapping groups of neurons peripherally and, due to the way the organism is wired up in the first place, any connections that have been made between one set and a response will also be effective for the set fired by the other stimulus. There are enough common cells fired by both stimuli to do this.

And then there are the induced similarities, such as the fact that the sound of the word "dog," the letters D—O—G, and the animal itself running along the floor, all in a sense can produce a similar kind of response. The response is sometimes similar and sometimes different, which means, I suppose, that you have overlapping sets of cells fired and sometimes the common part controls the response and sometimes the unique parts.

Now there is no reason why these groups should overlap in the animal's brain initially; that is to say, the sound "dog" may fire one group of neurons, and the image of the visual stimulus of the animal may fire a completely different set. Yet somehow or other they eventually come to give the same end result. The only explanation I can think of is that in the environment as it is set up, these things appear contiguously many, many times; so all you need is a mechanism of association between stimuli.

Now this association process can also work (and I think this is where the mathematical transformation theories of stimulus equivalence are not very adequate), it can also work for different aspects, or different views, or different translations in space of the same object, because these stimuli also appear contiguously very frequently. That is to say, if someone takes a book which is closed and he opens it, this is setting up a completely different stimulus pattern. But the two patterns, the back of the book and the inside of the book with its printing, appear contiguously so often, they have become associated with one another, so that now if you present either one you may get the same response.

CHAMMAH (*General Electric*): If for example you have random shapes of different size, at what point is a familiar pattern recognizable, such as a line or a square?

How close should the dots be to each other and what is the effect of area and so forth? Or to put it another way when does a human being start seeing patterns within a noise?

McCULLOCH: There are large numbers of measures of such things. It depends entirely upon the particular channel we are dealing with, whether your organism happens to have a good way of subtracting out noise, whether it can get a measure of the noise apart from the signal. If it can, then it can do a good job of finding the signal in the noise. It depends on whether it is looking for a pattern in the noise. This is an immensely complicated thing to answer and it depends on the particular sensory data and the organism or the machine.

LYNCH (*U.C.L.A.*): Long-term self-organizations, i.e. evolutionary processes in biology, versus relatively short-term self-organizations, i.e. cell division, self-programming computers, indicate the importance of time as a parameter, but in addition suggests that more and more of the elements in what might be called the free environment are being picked up (noise being fed on as Farley and von Foerster put it) and assimilated into the systems over the long term. It follows that the system must get more complex as time goes on and that the

elements in it become more and more specialized, that is to say, lose degrees of freedom. What implications does this have for the future of human social organizations?

NEWELL: I have only one reaction off the top on that one, which is that in some sense, if you take the power of learning these differences, I am not at all sure the proposition as stated holds, because the system is grabbing off information from the environment all the time and defining pairs of objects and finding out some new differences and building up the new tests. For a while the strings of tests that it may get may have larger and larger numbers of symbols. As it is forced to work with these it may try to find some new objects and in fact may attain a simple and neat set of differences. In some sense this program is continually getting better and better matched to the environment. So it is not at all clear that as it absorbs information from the environment its organization becomes more complex.

ABSTRACT OF A PAPER BY E. W. BASTIN (*King's College, Cambridge, England*)
CONCERNING A SELF-ORGANIZING MACHINE (CASPAR) HAVING A HIERARCHICAL
STRUCTURE OF ALGEBRAIC LEVELS

A SYSTEM of the "self-organizing" type is described which represents the logical structure of a scientific theory by means of *levels* arranged in a hierarchy of decreasing complexity. This model is particularly applied to physical theory, whose central concepts are assumed to derive their interpretation from the way we describe the behavior of a free particle in a field. The model differs from the standard model of a scientific theory in which hypotheses exist in levels in a hypothetico-deductive scheme and in which the meanings of hypotheses in the top levels are provided by the direct interpretation of the hypotheses at the lowest levels which are deducible from them. In the present model all levels are directly interpreted, but the lower (more complex) levels enable us to make a more detailed analysis of the experimental field.

A level is represented in the model by a square array of elements $a_{ij}(i, j = 1 \dots n^2)$ in which any element a_{ij} represents the probability of transition between two elements i, j in a vector $b_k(k = 1 \dots n^2)$. The elements of b are then formed into a square array of side length n , so that we now have a second, higher, level of elements $b_{ij}(i, j = 1 \dots n)$. It follows that at any level the elements change as the state of the system changes: these changes are determined by feedback from other elements in the level summed over a fairly long period so that the system has a memory and changes in the matrix are slow. Three levels are at present envisaged. The top level contains 4 components, the next 16, and the level below that, 256. This arrangement with a top level having four components corresponds to the inference in physics of a dimensional scheme, three space co-ordinates and one time co-ordinate, and the possibility of reducing this dimensionality to three under simple conditions (thus making the system nonrelativistic) is imposed as a constraint upon the system at the top level. The effects of this constraint at the various levels then make it possible to identify certain features of the working of the machine with more detailed physical concepts.

The hierarchy in the model may be thought of physically in the following way. The macroscopic properties of matter are usually supposed to be in principle calculable by a process of averaging over the individual effects due to the smaller systems that compose it. Moreover, these smaller systems are themselves assumed to arise by the averaging of smaller systems, and so on. One may therefore think of a hierarchy of *states of aggregation* of physical systems to each of which different sets of laws apply, and thus one may—in a natural way—link the successive stages of this aggregation process with the successive hierarchical levels in the CASPAR model. Now our most detailed physical theory—describing the least aggregated state of matter of which we have knowledge—is the theory of the elementary particles. We therefore seek to interpret the lowest level of our model in terms of elementary particles, and a provisional technique for accomplishing this interpretation, and thus of incorporating experimental data from the field of elementary particle physics into the model, is described.

The technique depends upon a *theory of clumps* being developed by I. J. Good, R. M. Needham and A. F. Parker-Rhodes from a device due to T. T. Tanimoto. Naturally occurring objective groupings (clumps) are progressively built up by an iterative procedure from experimental data presented in the form of a binary incidence table in which each of a set of varieties of experimentally discernible objects is specified according as it possesses, or does not possess, the individual members of a set of n properties. It was suggested by Margaret Masterman that a binary incidence table consisting of the empirically recognizable elementary particles specified in terms of a set of "quantum numbers" might be taken as one of the square arrays (levels) in the CASPAR model. Behind this suggestion lies the hypothesis that elementary particles, both known and unknown, should arise as the clumps in this array, and that the model might be used to discover these clumps.

The theory of clumps is valuable when no theoretical means of classification exists. This is the case with the newly discovered particles, except in so far as the theory of the interaction of the electron with the electromagnetic field acts as a general guide. Thus simple elaborations and extensions of the theory of the electron provide us with a set of properties (quantum numbers) in terms of which we can describe particles in an experimentally well-defined manner. Moreover, a number of experimentally distinguishable particles exist of each of which we can assert either that it has, or that it has not, the property specified by each quantum number. (Here I assume a slight change to have been made in current nomenclature to convert quantum numbers into binary properties so that, for example, the quantum number "charge" which can have three values, is taken as three distinct, single-valued, properties). The theory of clumps now enables us to discard the implicit assumption (already strongly suspect on account of the diversity of new particles) that a satisfactory system of mechanics ought to predict unambiguously the properties of the particles of which matter is made up so that concepts provided by theory have a natural correspondence to properties exhibited by particles. The clump theory allows a new freedom. It *may* happen, according to our approach, that some clumps are consistent with groupings determined by the laws of physics as we know them, whereas other clumps—equally probable *a priori*—are inconsistent with them. The paradigm case of

the former class is the electron: particles of the opposite class are unstable. The laws of physics are represented in the CASPAR model by the structure of the upper levels, together with a constraint imposed on one element at the 4-level to distinguish it from the rest and so incorporate the space-time distinction into the model.

This superimposed structure will cause some clumps of elements at the lowest level to produce stable modes of interaction of the model, whereas other clumps will be associated with unstable modes of interaction or oscillation. The basis of interpretation of the activity of the model therefore consists in identifying stable and unstable modes of activity with stable and unstable particles respectively, and the existence of such stable and unstable modes of behavior has already been suggested by the experimental work that has been described. It is hoped that further investigation of stabilities of clumps will show agreement with particles known to exist, and may predict associations or clumps of properties corresponding to as yet unknown particles. An algorithm for relating the similarities on which clump theory depends to particle masses is suggested as a first step in the practical use of the model.

A machine reproducing the top two levels of the model has been constructed in close collaboration with G. Pask using standard switching circuitry and servo-techniques, and the construction of this model has been of great assistance as a heuristic device throughout the early stages of the theory. This model is at present located in the Engineering Department of Swarthmore College, Swarthmore, Penn., U.S.A. where, with the advice and assistance of Dr. Carl Barus, parts of the machine have been replaced by simulation on a standard analog computer. A program representing these same two levels on the Cambridge digital computer, EDSAC, has been written and run by Needham. It is expected that future work will mainly utilize digital methods.

THE MECHANIZATION OF THOUGHT PROCESSES

Dr. A. M. UTTLEY

National Physical Laboratory, England

FIRST of all I must thank the organizers of this conference for the honor of being asked to speak at this evening banquet. I am in some doubt as to whether this should be an evening lecture or an after-dinner speech; unwisely I am going to attempt to make it both.

Firstly, I would like to survey the field of this conference and secondly, I would like to express some thoughts on its impact on the life and thought of the average man.

For the last ten years or so we have witnessed attempts to imitate and explain thought processes in terms of the physical sciences. There have been different approaches and different degrees of progress but there are slowly emerging a number of common and agreed ideas. I think it is now possible to put together into a single pattern some of these ideas, and I will tie them to a diagram which is so simple that no lantern slide is needed. A black box labelled "computer" has three inputs of which the first is from transducers which signal the changing state of the external world; the computer also has an output to controllers of the external world. The action taken by the controllers is reported back to form a second input to the computer. In parallel with this loop there is a second one; the consequences of the control actions in the external world appear at the first input.

There is a very important third input to the computer which signals the goal of the system.

The first input signals the state of the external world; this may be the state of a chemical process control plant or of a machine shop, a Russian manuscript awaiting translation, the contents of a library or other index, the present state of a mathematical process, sights

or sounds awaiting automatic recognition or indeed the same energy falling on the sense-receptors of an animal. In all these problems selection must be made before computation begins. In the first input channel we must insert a filter F and there is wide agreement that the mathematical principles of set theory must be embodied in this filter. One must select a class of inputs in terms of the presence or absence of properties. As the number of properties is increased the class narrows down and items irrelevant to the problem in hand are not passed on. The filtering is in terms of continuously varying probabilities determined in the computer, and I will return to this point later.

Turning now to the computer, we can distinguish two broad classes of processing, deductive and statistical. In consequence there emerge from the computer propositions which follow from the data and hypotheses of varying probability. The latter will arise in the analysis of a census return, or of the behavior of an industrial or a mathematical process, or in a man's observation of his muscular movements when attempting to acquire some new motor skill.

It is these computer outputs which are used to control the external physical world. Deductive outputs are used in existing servo-controllers because sufficient mathematical information has been fed in; a good example is the automatic pilot in an aeroplane. A very important new fact is that if there is insufficient information for this deductive control, there still can be control in terms of hypotheses arising in the computer. But there must be two new features in the system. Firstly there must be random trial of different forms of control; these will be initiated by some source of random noise. The computer will then receive information of:

- (a) the present state of the external world;
- (b) the control changes which have been tried;
- (c) the consequences in the external world.

Based on these facts the computer can compute the probabilities of hypotheses of the following form:

"In given circumstances a certain control will lead to particular consequences."

The second new feature is that a third input, introduced here for the first time, must select one of the consequences as a desirable goal. When this input occurs the computer can continually calculate, for each control movement previously tried, the probability that it

will lead to the goal; all these probabilities will continually alter with the changing state of the external world.

Automatic control in terms of hypotheses consists, then, in causing the above probabilities to modify the probability distribution of the noise source which generated the originally randomly chosen control movements; there must be an increased probability that the control which was successful in the past will be tried again. The principle of classification previously brought in for filtering input data must be used again in the selection of control changes; the same idea of selection by narrowing classes seems to be essential.

The principles here described appear to apply whether the goal is the optimization of a physical or a mathematical process. An example of the latter would be the problem of solving a set of simultaneous logical equations; if the number of variables is large all possible sets of values cannot be tried. Only if information is gained and used, during random trial, can such a problem be solved in a reasonable time.

Two further feedback loops must be added to the diagram. The filter which selects input data for relevance can act only on past experience, and it too, must be under the control of the hypotheses fed out from the computer; i.e. selection of input must be by the same principle on which random control is selected.

Lastly, the third "goal" input may be one of a series of sub-goals selected as above by an evolving master-goal within the computer.

The research covered by this conference raises a number of very important questions in the mind of the average man. Will there be unemployment? Shall we make brains? Are we debunking man?

As to the first question, machines will slowly take over from men all the tedious thinking which yet must be done without error; all the sorting and checking and counting and repetitive calculation which, frankly, makes machines of us. In less than a century such work will be as unthinkable as the transport of goods on the backs of slaves and man will be released for creative work which will satisfy him and give him joy. Rather than unemployment there will be re-employment in the "clerical revolution" which began ten years ago. But farsighted planning is essential.

Secondly, we shall not make brains any more than we shall make muscles. The aim, rather, is to understand the manifold functions of the brain and so ourselves. Instead of using the word "mind" we

prefer, to-day, to use the word "thinking." It embraces a large number of activities as yet little understood but already we may venture to resolve the "thinking-matter" problem by suggesting that thinking is a property of matter, matter in its most highly organized state—the nerve cell. And the word "consciousness" may, like "ether," disappear from our scientific language, not by denying obvious facts, but from a deeper understanding of them.

Using the languages of set theory and probability we may suggest that an idea is a probability relation between a set and a subset. Learning consists of the construction of supersets from subsets. From bricks we build houses; from houses we build towns. Whence come the bricks? The smallest bricks are our sensations, but from these alone a wolf-child will be constructed. Our parents and teachers give us much larger bricks; from these we build the hypotheses which not only determine our goals but also the filter with which we select further bricks. The hypotheses determine what I will and will not do in certain circumstances—and this is my character. Subsets of my character with its evolving goals pass on to those I meet, so I do not live in vain.
